1990

Applications of resampling methods to the estimation of ecological diversity

Clarice Azevedo de Luna Freire
Iowa State University

Order Number 9101350

Applications of resampling methods to the estimation of ecological diversity

Freire, Clarice Azevedo de Luna, Ph.D.

Iowa State University, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Applications of resampling methods to the estimation of ecological diversity

by

Clarice Azevedo de Luna Freire

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1990

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

Diversity Measures for Ecological Communities

There is an enormous variation among ecological communities with regard to their species composition and structure. This includes such features as the numbers of rare species and the numbers of prevalent or common species that are included in the concept of species diversity. Two aspects of species diversity that are commonly considered are:

- the number of species present in a community - *species richness*,
- the evenness of the relative abundances of the species - *equitability* or *evenness*.

Measures of diversity that focus on the number of species are generally referred to as measures of "species richness" instead of "species diversity". Measures of heterogeneity are more commonly called *diversity indices*. A number of diversity indices have been proposed to quantify various aspects of species diversity. Typically, a diversity index is a function of both the number of species and the relative abundances of the species. Hereafter, the number of species is denoted by S, and $\pi_1, ..., \pi_S$ are used to denote the relative abundance of species in the community. Measures of diversity generally satisfy the following two basic properties. The largest value of the index is obtained for a completely even community, i.e., $(\pi_1, ..., \pi_S) = (1/S, ..., 1/S)$.

For two completely even communities, the one with more species will have a largest value of the index. One measure that satisfies these properties is the Shannon (1949) index of diversity,

$$H' = -\sum_{i=1}^{S} \pi_i log(\pi_i). \tag{1.1}$$

The Shannon index is a member of a class of information theory functions called entropy measures. Simpson (1949) proposed

$$\lambda = \sum_{i=1}^{S} \pi_i^2$$

as a measure of the concentration of the species in a community. The Simpson index (also called the Gini index),

$$D = 1 - \sum_{i=1}^{S} \pi_i^2, \tag{1.2}$$

is a function of the concentration measure that also satisfies the two basic properties of a diversity measure.

Several authors have proposed classes of indices of diversity that are generalizations of H', λ and D. Examples are:

- $C_{m,n} = \sum_{i=1}^{S} \pi_i^m (-log\pi_i)^n$, for m,n=0,1,2,..., proposed by Good (1953),
- $\Delta_\beta = (1 - \sum_{i=1}^{S} \pi_i^{\beta+1})/\beta$, for $\beta \geq -1$, proposed by Patil and Taillie (1982).

A more complex approach for measuring diversity involves parametric modelling. Diversity is expressed as a function of the parameters of some probability distribution function or probability density.

Ecologists have expressed different opinions about measures of diversity. A strong criticism was voiced by Hurlbert (1971), particularly against the Shannon index and parametric approaches. He favored the Simpson index and other measures

directly related to the species abundances. Peet (1974) pointed out that with many potential diversity indices available, an investigator "should be able to select or design an index emphasizing that aspect of diversity he is most interested in measuring". Analyzing the contributions of individual species to the index, he concluded that the Shannon index is more sensitive to changes in rare species, and the Simpson index is more sensitive to changes in dominant species. In discussing the usefulness of Shannon's index as a diversity measure, Colinvaux (1978) stated that "it speaks to a very real difficulty we have in describing biological systems. The measure helps with our perennial problem of the common species and the rare, particularly the proper description of commonness and rarity.", "... it allows us to collapse estimates of species richness and species commonness into a single statement". Colinvaux strongly condemned the practice of using the index to measure features other than diversity, as for example, the stability of a community.

Relations between two measures of diversity (species richness or a diversity index), or between a measure of diversity and an environmental variable are also of major interest to ecologists. Two classical examples are a study on diversity of birds by MacArthur and Wilson (1961), and a study on diversity of lizards, by Pianka (1975). MacArthur and Wilson (1961) investigated the relation between the diversities of birds species and plant species, and the relation between the diversities of birds species and plant foliage-height (trees are classified according to number of layers of branches and density of foliage). Observing plots in deciduous forests of eastern United States, they found that bird species diversity increased with foliage-height diversity. A less strong correlation was detected between the two species diversity. Pianka (1975) analyzed communities of lizards in deserts areas of North America. He

found that lizard species richness was correlated to plant-volume diversity and not to plant species diversity. Fauth et al. (1989), in a study of communities of reptiles and amphibians in tropical forest in Costa Rica, examined the relation between species richness and elevation, and species richness and leaf-litter depth. They reported that species richness was negatively correlated with elevation and positively correlated with leaf-litter depth. Morton and Davidson (1989) studied harvest ant communities in Australian arid zones and compared their results to a previous study of harvest ants communities in North American deserts. They analyzed the relation between number of species and mean annual precipitation, and Shannon index of diversity and mean annual precipitation. Diversity measures were found to be linearly related to precipitation levels for communities in North America, but no relation was evident for the Australian communities.

Estimation of Ecological Diversity

There are many problems related to the estimation of ecological diversity. The definition of a sampling unit is a major one. In many cases, it is impractical to have the sampling units correspond to individual animals or plants. Random samples of areas (or volumes) often provide a reasonable alternative. Randomly sampling areas usually corresponds to selections of clusters of individuals. In the absence of some natural way to define areas, the dimensions and shapes for the sampling unit can be arbitrarily defined. These sampling units are called *quadrats* and in the case of land areas they are often specified as a grid of rectangles of identical size. Nilsson et al. (1988) used different types of quadrats for the estimation of the number of species of carabid beetles and the estimation of the number of species of land snails, in islands

at Lake Mälaren, Sweden. In the case of carabid beetles, each quadrat was a $10 \times 10m$ plot of land with nine pitfall traps arranged in a grid. In the case of land snails, each quadrat was a 0.1 m^2 plot of land, from which all litter, the uppermost soil layer and all herbs were removed for subsequent analysis. Even when there is a natural choice, some arbitrariness may prevail in the specification of quadrats. For example, Minshall et al. (1985) studied species richness of benthic invertebrate communities in streams. They sampled rocks from the bottom of the stream and removed the invertebrates adhering from them. These rocks are claimed to "behave in an analogous manner to oceanic islands". The choice made was to sample rocks with approximately 400 cm^2 of surface area.

Another procedure commonly adopted is to observe *transects*. These are contiguous plots of land, or volumes of water, etc. For example, Slobodkin et al. (1974) in a study about diversity of fish in coral reefs defined their transect as determined spaces, at every 10 meters along the reef wall. At each spot, all species seen were recorded, during an interval of 20 minutes. Transects are not analyzed in this study, although, if several transects are chosen randomly, they could be interpreted as quadrats.

The underrepresentation of rare species in a sample constitutes another problem. Rare species generally occupy proportionally small areas within the boundaries of a community, and are not as likely to be observed as more common species. The estimation of species richness tends to be more aggravated by the occurrence of rare species than the estimation of diversity (heterogeneity) indices.

Parametric and nonparametric approaches for the estimation of diversity are found in the literature. In a parametric approach some probability distribution is specified for species abundances (that is, the probabilities of observing a species with

r individuals, r=1,2,..., or the probability density if abundance is a continuous variable, such as area covered by plants or total weight of fish). Diversity is typically expressed as some function of the parameters of the distribution. This approach requires an assessment of whether the parametric model is appropriate for the data under consideration. Nonparametric approaches do not impose a model and simply use the number of observed species and the number of individuals observed for each species to estimate species richness or heterogeneity. The development of nonparametric methods to the estimation of ecological diversity is important because the enormous differences among types of biological communities and the arbitrary nature of quadrats makes it impossible to formulate a general parametric model suitable to every community. Resampling methods, such as the jackknife and the (nonparametric) bootstrap, can be applied as nonparametric estimation techniques that only require the selection of a sample of quadrats, based on some sampling design. Several approaches to the estimation of ecological diversity are considered in the following literature review.

Estimation of an index of diversity - α

An early attempt to devise and estimate a diversity measure was made by Fisher et al. (1943). They assumed that the expected frequencies of species with r individuals, in a random sample (of individuals), are $\alpha\ X^r/r$, r=1,2,..., for some $\alpha > 0$ and $0 < X < 1$. This is called the log-series distribution. Estimates of α and X are obtained by solving the equations $S^* = \sum_{r=1}^{\infty} \alpha\ X^r/r = -\alpha\ log(1-X)$ and $N^* = \sum_{r=1}^{\infty} \alpha\ X^r = \alpha\ X/(1-X)$, where N^* is the total number of individuals in a sample and S^* is the number of species in a sample. The mathematical relationship

between the number of species and the total number of individuals in a sample that follows from these two equations is

$$S^* = \alpha \, log(1 + N^*/\alpha) \ . \qquad (1.3)$$

For large samples, S^* can be approximated by $\alpha \, log(N^*/\alpha)$. Therefore, $S' - S'' \doteq \alpha \, [log(N'/\alpha) - log(N''/\alpha)] = \alpha \, log(N'/N'')$, where S' and S'' are the number of species observed in samples of sizes N' and N'', respectively. Thus, the parameter α can be interpreted as the increase in the observed number of species due to some increase in the sample size. The relationship (1.3) implies that the number of species increases with sample size. On the contrary, α was shown to stabilize as sample size increased. Consequently, α was chosen as a measure of the diversity of the community, and it was labelled by Fisher et al. (1943) as an *index of diversity.* Fisher et al. (1943) provided an approximation for the variance of $\hat{\alpha}$, making possible comparisons among communities. This model was fit to extensive data supplied by collectors of butterflies and to a large body of data on insects captured by means of light-traps. Fisher et al. (1943) did not approach the problem of estimation of the number of species of an entire community. Whenever number of species was considered, it was considered conditionally on the observed sample.

The fit of the log-series model to data sets from a variety of ecological communities was not always adequate. Krebs (1978, chapter 23) points out that "The logarithmic series implies that the greatest number of species has minimal abundance, that the number of species represented by a single specimen is always maximal". Krebs (1978) cites examples of published data where species abundances do not seem to follow a logarithmic series pattern. In these communities, the majority of the observed species has some average number of individuals sampled, and few species

have either large or very small number of individuals in the sample. Therefore, other distributions were considered to model species abundance.

Pielou (1975, chapter 3) reviews parametric procedures for the estimation of the number of species in a community, when the truncated negative binomial and the lognormal distributions are used as species abundance distributions. In both cases the estimator for the number of species in the population is the observed number of species divided by 1-p, where p is the estimated probability of not observing a species. Procedures for estimating the variances of these estimators, however, have not been developed.

Parametric procedures require an assessment of whether the proposed model is appropriate for the data. Usually, accurate assessments can not be made with small samples. Moreover, as observed by Pielou (1977, page 292), the fit of a species abundance distribution may be impossible "if there are only a few species in a sample and each is represented by a different number of individuals". Furthermore, the approach of Fisher et al. (1943) is based on a simple random sample of individuals, which is very difficult to obtain in many situations.

Jackknife methods

Suppose that a quantity θ is estimated by $\hat{\theta}$, based on information from a single random sample of size n. By deleting elements one at a time, from the original sample (size n), n samples of size n-1 are created. Let $\hat{\theta}$ be the estimate derived from the sample of size n, and $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ the estimates derived from the created samples.

The jackknife estimator of θ defined by

$$\hat{\theta}_{JK} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i,$$

where $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_i, i = 1,\ldots,n$, often provides an improvement in the sense that $\hat{\theta}_{JK}$ has smaller bias than $\hat{\theta}$. This method also provides an estimate for the variance of the jackknife estimator,

$$\widehat{Var}(\hat{\theta}_{JK}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_i - \hat{\theta}_{JK})^2,$$

which might be used, in most cases, for the construction of confidence intervals for θ, using Student's t distribution. Basic information on jackknife methods is found in Miller (1974), Efron (1982), and Efron and Gong (1983).

Adaptations of the jackknife estimation to other sampling designs have been presented in the literature. Rao and Wu (1985) summarized several jackknife estimators based on stratified simple random sampling. Heltshe and Forrester (1983, 1985) applied jackknife methods in the estimation of diversity based on a random sample of quadrats, which can be viewed as an adaptation to one-stage cluster sampling. When the estimation is based on a random sample of n quadrats, the artificial jackknife samples are the n sets of n-1 quadrats created by successively deleting one quadrat from the full set of n quadrats. Heltshe and Forrester (1983) developed closed form expressions for the jackknife estimator of the number of species and for the variance of the estimator. The jackknife estimator tends to underestimate the number of species when there are many rare species in the community. In the second paper, Heltshe and Forrester (1985) examined the behavior of jackknife estimators for the Brillouin index and the Simpson index. For the Simpson index, the jackknife

estimator is unbiased, but the proposed variance estimator tends to overestimate the size of the variance for large samples.

Bootstrap methods

Once a simple random sample of n elements has been observed, consider all sets that can be constructed by selecting randomly, with replacement, n elements from the original sample. These sets are called *bootstrap samples.* Suppose that a quantity θ is estimated by $\hat{\theta}$. Each bootstrap sample provides an estimate of θ. Basically, a bootstrap method consists in analyzing the behavior of the estimator $\hat{\theta}$, as a discrete random variable whose support is the set of the estimates obtained from the possible bootstrap samples, with associated probabilities equal to their relative frequencies. Features such as expectation, variance, etc..., of the estimator $\hat{\theta}$ are estimated by the expectation, variance, etc..., of the value of the estimator for the possible bootstrap samples. Either explicit expressions for the expectation, variance, etc.,$\cdots$, are developed or approximations are based on a large number, B, of randomly selected bootstrap samples. In particular, estimates for the expectation and the variance of a bootstrap estimator can be approximated, respectively, by the average, $\hat{\theta}_{BT}$, and the sample variance, $(\hat{\sigma}_{BT})^2$, of B estimates from B bootstrap samples.

Confidence intervals for θ can be constructed in several ways, using bootstrap methods. A confidence interval might be centered on $\hat{\theta}$ or it can be centered around some bootstrap estimator for θ, such as $\hat{\theta}_{BT}$. An approximate $(1-\alpha)100\%$ confidence interval for θ is

$$\left[\hat{\theta}_{BT} - C_{\frac{\alpha}{2}}\hat{\sigma}_{BT} \quad , \quad \hat{\theta}_{BT} + C_{\frac{\alpha}{2}}\hat{\sigma}_{BT}\right],$$

where C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution. These

confidence intervals rely on the limiting (large n) normality of the estimators. For relatively small samples, there are other ways of using the bootstrap procedure to construct confidence intervals for θ, which are often more appropriate. The estimates obtained from B bootstrap samples, provide some information about the estimator's distribution. Their histogram might suggest, for instance, that the distribution is skewed. One way to use this information is to use the $(\frac{\alpha}{2})100^{th}$ percentile and the $(1-\frac{\alpha}{2})100^{th}$ percentile of the set of B estimates from the bootstrap samples as the lower and upper limits of an $(1-\alpha)100\%$ confidence interval for θ. This is called the " *percentile method* ". It does not require the specification of a parametric model for the distribution of population values.

Bootstrap methods are very flexible and have been adapted to a wide variety of applications; examples can be found in Efron (1982) and in Efron and Gong (1983). Bickel and Freedman (1984) analyzed the asymptotic behavior of a bootstrap estimator for linear combinations of means under stratified sampling design. It was concluded that the usual Lindeberg conditions guarantee that the bootstrap approximation of the t-statistic converges in law to the standard normal distribution. For the i^{th} stratum, the variance of the bootstrap estimator for the mean, under bootstrap sample scheme, is $(n_i-1)s_i^2/n_i^2$ and the variance of the sample average is estimated by s_i^2/n_i, where n_i and s_i^2 are the sample size and sample variance, respectively. Therefore, they conclude that some scaling is necessary, to use the percentile method. Rao and Wu (1988) presented extensions of bootstrap methods to some sampling designs involving clustering and stratification. This work is discussed in detail in later section.

In the extension of bootstrap methods to quadrat sampling, bootstrap samples

are created by selecting with replacement n quadrats from the original sample of n quadrats. Smith and Van Belle (1984), developed a bootstrap estimator for the bias of the observed number of species, $\widehat{S}$, in the estimation of the number of species, when quadrat sampling is employed. A bootstrap estimator for the number of species was defined as the observed number of species minus the bootstrap estimate of bias. Theoretical comparisons were made between the unconditional expectation of their bootstrap estimator of the bias of $\widehat{S}$ and the corresponding bias correction provided by the jackknife estimator presented in Heltshe and Forrester (1983). Those comparisons were based on a probability model that assumed random distribution of individuals in the community, and sampling of individuals from quadrats. In the presence of many rare species, both estimators were downward biased for small samples. Jackknife estimation was better for smaller samples, but for larger samples bootstrap estimation was better, since the jackknife method was overestimating the number of species. Smith and Van Belle (1984) did not analyze confidence intervals and no simulations were performed.

Pielou's estimation of Shannon's index

Pielou was interested in obtaining an estimator for Shannon's index and an estimator for its variance. Pielou (1975) described a procedure for the estimation of Shannon's index of diversity (1.1), based on a large random sample of n quadrats. It involves the calculation of the Brillouin index

$$H = \frac{1}{N} log \left(\frac{N!}{\Pi_{i=1}^{S} N_i!} \right),$$

where N_i is the total number of individuals for species i, i=1,...,S, and $N = \sum_{i=1}^{S} N_i$. First the quadrats are ordered randomly. Then, h_k values are calculated and plotted versus k, k=1,...,n, where

$$h_k = \frac{M_k H_k - M_{k-1} H_{k-1}}{M_k - M_{k-1}},$$

where M_k is the total number of individuals in the pool of the first k quadrats, and H_k is the Brillouin index calculated for the pool of the first k quadrats. It is argued that if the plot of h_k versus k shows an initial increase, and then becomes stable after some value k=t, the values of h_k, for $k \geq t$, are reasonable estimates of the Shannon index. Then, an estimator for the Shannon index and its variance are given by

$$\widetilde{H'} = \frac{1}{n-t} \sum_{k=t+1}^{n} h_k \quad \text{and} \quad \widetilde{Var}(\widetilde{H'}) = \frac{1}{n(n-1)} \left(\sum_{k=t+1}^{n} h_k^2 - n\widetilde{H'^2} \right),$$

respectively. Pielou reported that some researchers apply this procedure to several random orderings of quadrats and use the median of the estimates $\widetilde{H'}$ as the final estimate. No further analysis of this method was made.

A Bayesian estimator for the expected number of species

In the Bayesian approach the number of species, S, existing in a community is considered as a random variable. An empirical Bayes estimator for the expected number of species, E(S), was developed by Mingoti (1989), based on a sample of n quadrats. Her Bayesian model requires several assumptions and the definition of several random variables:

1. "For each species s_i, let p_i be the probability that species s_i is observed in a typical quadrat of the region. So is the same for any quadrat, $0 < p_i < 1$, for $i =$

$1, ..., S$.". The probabilities $p_1, ..., p_S$ are treated as independent and identically distributed random variables, with some continuous distribution function g(.) on the interval [0,1]. Mingoti considered a beta distribution;

2. X_i is defined as the "number of quadrats in the sample where the species s_i was observed", for i=1,..., S; "given p_i, $0 < p_i < 1$, X_i is a binomial random variable with parameters n and p_i...";

3. n_x is defined as the "number of species observed in exactly x quadrats in the sample", for x=0,1,..., n. To establish the distribution of n_x the following probability is defined: "For each species s_i the probability that it will be observed in exactly x quadrats in the random sample of n quadrats is given by

$$\gamma_x = \text{"...“} = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} g(p)\, dp \text{ ",}$$

for i=1,..., S and x=0,1,..., n. "Given S and γ_x, n_x is a binomial random variable with parameters S and γ_x"; therefore, the expected value of n_x, given S and γ_x, is $E_S(n_x) = S\gamma_x$, for x=0,1,..., n;

4. Z_m is defined as the "number of species observed in the second sample of m quadrats, which were not observed in the first sample of n quadrats", for $m \geq 1$. To establish the distribution of Z_m the following probability is defined: "for any species s_i the probability that it will be observed in the additional sample of m quadrats and it was not observed in the first sample of n quadrats is given by

$$\gamma^* = \int_0^1 (1-p)^n (1-(1-p)^m) g(p)\, dp \text{ .";}$$

"before the first sample is observed, Z_m has binomial distribution with parameters S and γ^*," and the expected value of Z_m, "before the first sample is observed", given S and γ^*, is $E_S(Z_m) = S\gamma^*$.

The final estimator is expressed as a sum of two estimators: the observed number of species, $\widehat{S}$, estimating the expected number of species for the sampled area, and an estimator for the expected value of Z_{N-n}, "before the first sample is observed" and given S.

By the above assumptions, $E_S(Z_{N-n}) = S\gamma^*_{N-n}$, before the first sample is observed, and $E_S(n_1) = S\gamma_1$. The random variable S, in the expression of $E_S(Z_{N-n})$, is arbitrarily replaced by $E_S(n_1)/\gamma_1$. No justification was given for using n_1 rather than n_2, or n_3, ..., or n_n, or possibly, an average based on all n random variables. Therefore, $E_S(Z_{N-n}) = [E_S(N_1)/\gamma_1]\ \gamma^*_{N-n}$ is estimated by $[n_1/\gamma_1]\ \gamma^*_{N-n}$, where n_1 is the observed number of species in only one quadrat (same notation as the random variable n_1).

Considering that the probabilities p_i, for i=1,...,S, are distributed according to a beta distribution with parameters α and β, γ_1 and γ^*_{N-n} are expressed as functions of the parameters α and β. The values of the parameters for the beta distribution function are typically unknown. To estimate those parameters, Mingoti assumes that for every observed species the probability of being observed in exactly x quadrats is

$$qx = \frac{\gamma x}{\sum_{x=1}^{n} \gamma x},$$

and, in terms of α and β,

$$qx = \frac{\binom{n}{x}\,\Gamma(x+\alpha)\Gamma(n+\beta-x)}{\sum_{x=1}^{n}\binom{n}{x}\,\Gamma(x+\alpha)\Gamma(n+\beta-x)}.$$

The random vector $(n_1, ..., n_n)$ is assumed to be distributed according to a multinomial distribution with parameters $(q_1, ..., q_n)$, given the observed number of species in the sample of n quadrats. The likelihood function of this distribution is

$$f(n_1, ..., n_n) = \frac{\hat{S}!}{\prod_{x=1}^{n} n_x!} \prod_{x=1}^{n} q_x^{n_x}.$$

Estimates for the parameters of the beta distribution are obtained by maximizing this likelihood function with respect to α and β. Such estimates do not have a closed form expression and some numerical procedure must be used.

The resulting estimator for the expected number of species is

$$\hat{S}_{Bayes} = \hat{S} + \frac{n_1}{n\,\hat{\alpha}}(n + \hat{\beta} - 1)\left[1 - \frac{\Gamma(n + \hat{\alpha} + \hat{\beta})\,\Gamma(N + \hat{\beta})}{\Gamma(n + \hat{\beta})\,\Gamma(N + \hat{\alpha} + \hat{\beta})}\right],$$

where

N=total number of quadrats,

$\hat{\alpha}$=estimate of first parameter of the beta distribution function,

$\hat{\beta}$=estimate of second parameter of the beta distribution function,

$\Gamma(x)$=gamma function evaluated at x.

The assumptions underlying Mingoti's model and estimation procedures are not likely to be well met by biological communities. Individuals tend not to be "evenly" distributed across quadrats as Mingoti assumes. The concept of "a typical quadrat" ("let p_i be the probability that species s_i is observed in a typical quadrat of the region") may be quite unreasonable. Biological communities often display spatial patchiness of individuals, with some species more abundant in particular areas than others. Some quadrats may contain high concentration of a particular species and

others may not. Generally there is a higher probability of observing a species in a quadrat with a high concentration of individuals from the species.

The Bayesian estimator above is a function of N (total number of quadrats). This might constitute a practical problem in applications to biological communities, since the exact value of N is rarely known (the exact definition of boundaries for a biological community is seldom attained). In some instances, the sampled proportion is very small, consequently, N is a very large value; for example, core samples (very small volumes), are frequently extracted from bottoms of lakes, rivers, etc..., in studies of benthic communities. Therefore, it would be necessary to investigate the behavior of the estimator, when the exact value of N is unknown and accurate estimates of N are not available.

Estimating Diversity under Complex Sampling

The quadrat sampling procedures described in the previous sections were based on a random selection of quadrats. In many instances, however, stratification and several stages of selections are employed, when sampling ecological communities. The following references are examples of complex sampling in ecological surveys:

- Ross et al. (1985) studied species richness in stream fish assemblages. To survey a stream, collections of fish were caught with a meshed seine (a large fishing net that has sinkers at the lower edge and floats at the upper). The stratification corresponds to microhabitats and "an effort was made to sample all available microhabitats (e.g., pools, runs, riffles)";

- Nilsson et al. (1988) studied species richness of carabid beetles in Swedish islands. Vegetation type defined six strata: deciduous forest, conifer-dominated forest, alderwood, shore meadow, grass meadow and rocky ground. Initially, $100m^2$ plots were selected. Subsequently, $0.1m^2$ subsamples plots, were taken from the larger plots;

- Morton and Davidson (1988) studied harvest ant community in Australian arid zones. Stratification was defined by vegetation formation. Most of the selected sites were chosen from two major vegetation formations, acacia shrublands and hummock grasslands, since they occupy large proportion of the arid zone, but "four other sites of diverse vegetation were also included...";

- In a study on diversity of birds in the city of Ames, Iowa, performed by James Dinsmore, Georgia Bryan and Bret Giesler, the strata are types of areas in the city: commercial, new residential, old residential, parks and green belts. From an initial sample of large areas, circular areas of approximate radius of $25m$ are surveyed.

Variances for estimators of diversity measures must account for the sampling procedure. The mathematical development of variance formulas is avoided with the use of resampling methods. In this section, only bootstrap methods are considered.

Although the definition of a bootstrap sample is dependent on the original sampling scheme, bootstrap methods for various complex sampling schemes share some common basic features. In general, a bootstrap sample is created by sampling units from the original sample, with replacement, following the same design used when sampling from the population. An estimate from a bootstrap sample is calculated in

the same manner that an estimate is calculated from the original sample.

In the extension of bootstrap methods to stratified simple random sampling, for example, a bootstrap sample is defined as a set of independent random selections, each performed with replacement, with one taken from each stratum involved in the original sample. Within each stratum, the bootstrap sample size is equal to the original stratum sample size. For a two-stage cluster procedure, in which clusters are selected at a first stage and simple random samples are selected from those clusters at the second stage, a bootstrap sample also requires a two-stage procedure selection. Initially, random selections of clusters, with replacement, are made from the set of clusters observed at the first stage of sampling. Then, units are randomly selected, with replacement, from each cluster chosen at the first stage of the bootstrap sampling. The number of units selected from a resampled cluster is the same as the number of units selected from that cluster for the original sample.

The estimation of a diversity index is a particular case of the general problem of estimating a function of a multivariate population mean. In the problem of estimating a function of a multivariate population mean, $f(\mu_1, ..., \mu_S)$, usually, estimates of the means are obtained, $\hat{\mu}_1, ..., \hat{\mu}_S$, and f is estimated by $\hat{f} = f(\hat{\mu}_1, ..., \hat{\mu}_S)$. In most cases, a closed form expression for the variance of $\hat{f}$ does not exist. In many situations, the Delta Method is useful in providing an approximation for the variance of $\hat{f}$,

$$Var(\hat{f}) \doteq \sum_{i=1}^{S} \sum_{j=1}^{S} \widehat{Cov}\left[\hat{\mu}_i, \hat{\mu}_j\right] f_i'(\hat{\mu}) f_j'(\hat{\mu}),$$

where $f_i'(\hat{\mu})$ is the first partial derivative of f with respect to the i^{th} coordinate, evaluated at the point $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_S)$.

Alternatively, in the bootstrap estimation of f, based on $\hat{f}$, individuals means are estimated from a bootstrap sample, $\hat{\mu}_1^*, ..., \hat{\mu}_S^*$, and f is estimated by $f^* = f(\hat{\mu}_1^*, ..., \hat{\mu}_S^*)$. The variance of $\hat{f}$ is estimated by the sample variance of a set of estimates f^*, from a large number of bootstrap samples. Furthermore, the sample percentiles for a large number of bootstrap values f^* can be used to approximate percentiles of the distribution of $\hat{f}$. There are a number of ways of using the bootstrapped percentiles to construct a confidence interval for f.

Modified bootstrap methods

Rao and Wu (1988) analyzed bootstrap estimators for functions of a population mean, based on the two sampling designs described in the previous section. The basic technique employed by Rao and Wu to improve a bootstrap estimator is described for the stratified design.

For the stratified simple random sample (with replacement) design, with L strata, the usual estimator of a population mean, $\overline{Y}$, is

$$\overline{y} = \sum_{h=1}^{L} W_h \, \overline{y}_h, \tag{1.4}$$

where W_h is a known weight corresponding to the relative size of the h^{th} stratum and $\overline{y}_h$ is the sample average for the original sample from the h^{th} stratum, h=1,...,L. The usual estimator of $f(\overline{Y})$ is $f(\overline{y})$. Suppose that n_h is the size of the original simple random sample selected from the h^{th} stratum, h=1,...,L, then a bootstrap is created by using simple random sampling with replacement to select a sample of size n_h from the original sample in the h^{th} stratum, h=1,...,L. The bootstrap estimator

of $\overline{Y}$, based on (1.4), for a particular bootstrap sample, is

$$\overline{y}^* = \sum_{h=1}^{L} W_h \, \overline{y}_h^*, \tag{1.5}$$

where $\overline{y}_h^*$ is the sample average in the bootstrap sample from the h^{th} stratum sample, for h=1,...,L. The bootstrap estimator of $f(\overline{Y})$, based on $f(\overline{y})$, for a particular bootstrap sample, is $f(\overline{y}^*)$. Conditional on the original sample, the variance of (1.5) is

$$Var_*(\overline{y}^*) = \sum_{h=1}^{L} W_h^2 \left(\frac{n_h - 1}{n_h}\right) \frac{s_h^2}{n_h}, \tag{1.6}$$

where s_h^2 is the sample variance of the observed sample from the h^{th} stratum, h=1,...,L. An unbiased estimator for the variance of (1.4) is

$$\widehat{Var}(\overline{y}) = \sum_{h=1}^{L} W_h^2 \, \frac{s_h^2}{n_h}. \tag{1.7}$$

Therefore, the variance of the bootstrap estimator is a consistent estimator for $Var(\overline{y})$, since the ratio of (1.6) to (1.7) converges to 1 as every stratum sample size increases.

Rao and Wu (1988) were interested in the case of bounded stratum sample sizes. It was argued that in this case the variance of the bootstrap estimator is not consistent in the estimation of $Var(\overline{y})$. A modified version was defined where the bootstrap sample sizes for the strata were arbitrarily chosen. From each bootstrap sample, the bootstrap estimator for $\overline{Y}$, based on $\overline{y}$, is defined as

$$\tilde{\overline{y}} = \sum_{h=1}^{L} W_h^2 \left[\overline{y}_h + \left(\frac{m_h - 1}{n_h}\right)^{1/2} (\overline{y}_h^* - \overline{y}_h)\right],$$

where m_h is the size of the created sample from the observed sample from the h^{th} stratum, h=1,...,L. It can be shown that $Var_*(\tilde{\overline{y}}) = \widehat{Var}(\overline{y})$. The modified bootstrap

estimator for $f(\overline{Y})$, based on $f(\overline{y})$, for a specific bootstrap sample, is $f(\tilde{\overline{y}})$. Rao and Wu compared $Var_*[f(\tilde{\overline{y}})]$ with the estimate for the variance of $f(\overline{y})$ obtained by the Delta Method. Since these estimates differed by a random variable $O_p(n^{-2})$, where n is the total sample size, they concluded that $Var_*[f(\tilde{\overline{y}})]$ is consistent in estimating $Var[f(\overline{y})]$. The authors suggested the use of $m_h = n_h - 3$, whenever $n_h \geq 5$, and advised against using $m_h \geq n_h$, saying it could lead to negative estimates even if the parameter of interest is positive. A correction is incorporated to the modified bootstrap version, when sampling from the population was made without replacement.

In a two-stage cluster design, n clusters are randomly selected from a set of N clusters. A simple random sample of size m_i is taken from the i^{th} cluster, which has M_i elements, if that cluster is selected at the first stage (i=1,...,N). The population mean by element (see Cochran, 1977, pages 249-250) is

$$\overline{\overline{Y}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}}{\sum_{i=1}^{N} M_i},$$

where Y_{ij} is the j^{th} element of the i^{th} cluster, j=1,$\cdots$,M_i, i=1,$\cdots$,N. An unbiased estimator for $\overline{\overline{Y}}$ is

$$\overline{\overline{y}} = \frac{1}{M} N \left\{ \frac{1}{n} \sum_{i=1}^{n} M_i \left[\frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \right] \right\},$$

where $M = \sum_{i=1}^{N} M_i$ and y_{ij} is the j^{th} selection from the i^{th} selected cluster, j=1,$\cdots$,m_i, i=1,$\cdots$,n.

A bootstrap sample is created by randomly selecting with replacement n clusters from the original sample of n clusters. A random sample with replacement of size m_i is selected from the i^{th} cluster, whenever the i^{th} cluster is selected for the bootstrap

sample. For example, if cluster 2 is selected three times for a bootstrap sample, three independent samples of size m_2 are drawn with replacement from the original sample from cluster 2. Rao and Wu (1988) proposed a modified version for a bootstrap estimator for $f(\overline{\overline{Y}})$, based on $f(\overline{\overline{y}})$. For each bootstrap sample, an estimate for $\overline{\overline{Y}}$ is

$$\begin{aligned}\tilde{\overline{\overline{y}}} &= \overline{\overline{y}} + \frac{1}{n}\left[(1-f_1)\frac{n}{n-1}\right]^{1/2}\sum_{i=1}^{n}\left[\frac{1}{M}N(M_i^*\overline{y}^*) - \overline{\overline{y}}\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\left(f_1(1-f_{2i}^*)\frac{m_i^*}{m_i^*-1}\right)^{1/2}\right]\left[\frac{1}{M}N(M_i^*\overline{y}^{**}) - \frac{1}{M}N(M_i^*\overline{y}^*)\right]\right\},\end{aligned}$$

where, $f_1 = n/N$, $f_{2i} = m_i/M_i$, m_i^*, M_i^* and $\overline{y}_i^*$ are, respectively, the values for m_i, M_i and $(1/m_i)(\sum_{j=1}^{m_i} y_{ij})$ for the i^{th} resampled cluster, $f_{2i}^* = m_i^*/M_i^*$, and $\overline{y}_i^{**} = (1/m_i^*)(\sum_{j=1}^{m_i^*} y_{ij}^{**})$, where y_{ij}^{**} is the j^{th} random selection with replacement from the i^{th} resampled cluster.

The variance, under bootstrap sampling scheme, of the estimator $\tilde{\overline{\overline{y}}}$ is equal to an unbiased estimate for the variance of $\overline{\overline{y}}$. The dependency on the total number of elements in the population, M, and on the total number of individuals, M_i, in the observed clusters is a detrimental feature of this estimator, since these quantities are often unknown. A technique to eliminate the quantity M from the expression has been suggested, but the M_i values still need to be known. The number of units selected from a resampled cluster is kept the same as the number of units selected from that cluster for the original sample. No investigation was made to verify if the numbers of selections of resampled clusters and within resampled clusters should be different.

Further developments on complex sampling

In the subsequent chapters, additional bootstrap methods for the estimation of functions of a population mean are derived under cluster sampling schemes, where at the last stage chosen clusters are thoroughly surveyed. Cluster sampling is relevant because most ecological community surveys can not be based on individual sampling. Results can be extended to cluster sampling with additional stage of subsampling, where at each stage clusters are subdivided into the same number of smaller clusters, and the clusters selected at the final stage of sampling are completely observed.

Bootstrap estimators for $f(\overline{\overline{Y}})$ are based on $f(\overline{\overline{y}})$, where $\overline{\overline{y}}$ is a biased estimator for the population mean,

$$\overline{\overline{y}} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij} \; / \; \sum_{i=1}^{n} m_i,$$

which is not a function of the total number of elements in the population. Modifications are developed, following the technique of Rao and Wu (1988), to account for finite population situation and sampling without replacement. Bias-corrected versions of bootstrap estimators are developed. Although emphasis is given to the estimation of functions of proportions, which characterize a diversity index, the results can be applied to other functions of means of environmental variables used in measuring diversity.

Simulation designs for estimation of diversity

Simulation studies are used to compare properties of various estimators in subsequent chapters. Communities are randomly established on unit squares, by randomly generating the numbers and locations of patches of individuals. The resulting com-

munities are analyzed using samples with different quadrat sizes and different sample sizes. These procedures are similar to those used by Heltshe and Forrester (1983).

In the Chapter on estimation of species richness, communities containing 25 species were formed. The locations and numbers of "parents" and "offspring" for each species were determined in the following way:

- the number of parents and the number of offspring per parent followed two independent Poisson distributions whose parameters were pre-determined; the number of offspring per parent followed the same Poisson distribution for all parents within a species (these values are given in table (1.1));
- the location of a parent was determined according to the uniform distribution on the unit square;
- the location of each offspring was determined by two random variables. The radius of a circle centered on the parent location was determined by the absolute value of a normal $(0, \sigma^2)$ random variable and the sign of the normal random variable defined the location of the offspring to be in either the upper or lower semicircle. The uniform $[0, 180]$ random variable determined the exact location of the offspring on a semicircle, by providing the angle between the radius and the horizontal axis. The value of σ^2 was constant to all species within a community.

Four communities, labelled A, B, C and D, were assembled. For A and B the value of σ was 0.14, as in the work of Heltshe and Forrester (1983), and $\sigma = 0.02$ was used for C and D. According to the normal distribution, approximately 95% of the offspring (belonging to a same parent) are expected to be located within a circle of

Table 1.1: Poisson parameters for the number of parents and offspring

Community A		Communities B and C		Community D	
parents	offspring	parents	offspring	parents	offspring
30	10	100	14	15	2
30	10	80	10	15	2
30	10	40	7	15	2
30	10	40	7	15	2
30	10	40	7	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	2	15	2
10	10	15	1	15	2
10	10	15	1	15	2
8	8	15	1	15	2
8	8	15	1	15	2
8	8	15	1	15	2
8	8	15	1	15	2
5	5	15	1	15	2
5	5	15	1	15	2
5	5	15	1	15	2
5	5	15	1	15	2

radius 2σ, i.e., within a circle of area $\pi(2\sigma)^2$. Therefore, the choice of $\sigma = 0.14$ would produce patches of individuals covering large areas; for example, 40% and 54% of the patches with 10 offspring and 15 offspring, respectively, would have area greater than 0.25. Alternatively, with $\sigma = 0.02$, 40% and 54% of the patches with 10 offspring and 15 offspring, respectively, would have area greater than 0.005. The differences in dispersion and numbers of individuals resulted in different levels of difficulty for the estimation of the number of species from communities A to D. The estimation of the number of species is least difficult for for community A, designed to have many offspring per parent; communities B and C differ only in the dispersion of the offspring around parents and most of the individuals in these communities belong to one of five species. Therefore, B and C represent very uneven communities. community D contains mostly rare species with small dispersion among individuals in a group.

The FORTRAN program written to generate the communities is presented in Appendix A.

To compare the effects of using different numbers and sizes of quadrats, sample sizes (i.e., number of quadrats) used in the simulation studies were chosen to allow either 2.5% , 5% and 10% of total area to be included in the sample. The quadrat sizes considered were:

small quadrats - ▫ - unit square divided into 784 quadrats (28×28);
medium quadrats- □ - unit square divided into 400 quadrats (20×20);
large quadrats - ☐ - unit square divided into 196 quadrats (14×14).

For each one of the 9 sampling schemes presented in Table 1.2, one hundred samples were selected, from each community. Point estimates and 90% confidence intervals for S, the number of species, were calculated. The Bayesian estimator described in

Table 1.2: Sampling schemes - number of quadrats and quadrat size

Total Sampled Area		
2.5%	5%	10%
20 ▫	40 ▫	80 ▫
10□	20□	40□
5 ☐	10 ☐	20 ☐

notation	
small quadrat	▫
medium quadrat	□
large quadrat	☐

previous section was also evaluated for the first 20 samples of large quadrats selected from each community.

For the estimation of diversity indices one community containing 10 species was generated on a unit square. The procedure used was similar to the one performed by Heltshe and Forrester (1983), where "parents" are located at random, and "offspring" are located around each parent. The numbers of parents, the number of offspring, and the location of offspring were determined as in Heltshe and Forrester (1983). The locations of parents were made differently, in order to introduce dependency among some species. The locations of parents for species 1 (the species with the largest number of parents) were randomly assigned according to the uniform distribution on the unit square. For species 2 and 3, the locations of parents were determined by adding random variables distributed as uniform [-0.1,0.1] to the coordinates of

parents of species 1, until the specific numbers of parents of species 2 and 3 were reached. Similarly, species 4, 5 and 6 were generated. The locations of parents for species 7 were determined according to the uniform distribution on the unit square. To locate a parent of species 8, a random variable following the uniform distribution on the unit square was initially observed; it was used if that point was not located within circles of radius 0.1 centered at the locations of parents of the first species, otherwise, another random variable was generated. Similarly, species 9 and 10 were generated. Therefore, two sets of species display positive association, {1,2,3} and {4,5,6}, and two sets of species display negative association, {7,8} and {9,10}. Three values were given for the dispersion parameter: for species 1, $\sigma = 0.05$, for species 3, $\sigma = 0.01$ and, for all other species, $\sigma = 0.02$. This community is characterized by dominance of species 1,2 and 3, and by small dispersion among individuals within a species.

For the sampling process, the unit square was subdivided into square quadrats of equal area. Three quadrat sizes were used, corresponding to grids of 14×14, 20×20 and 28×28 quadrats. The sample sizes considered correspond to sampling 2.5% , 5% and 10% of the unit square. For each one of the nine sampling schemes, defined by quadrat size and percentage of area sampled, one hundred independent samples of quadrats were selected from the community.

For each one of the nine sampling schemes presented in Table 1.2, one hundred samples were selected from the community. Point estimates and 90% confidence intervals for the Shannon index (1.1) and the Simpson index (1.2) were calculated, using several definitions of resampling estimators.

Explanation of Dissertation Format

This dissertation follows the alternate format, where individual papers are presented. In the first paper (Chapter 2) applications of resampling methods to the estimation of species richness, with basis on a random sample of quadrats, are investigated through simulations. A bootstrap estimator for the variance of the jackknife estimator is proposed. An illustration with real data is presented. The second paper (Chapter 3) focusses on the bootstrap estimation of functions of proportions under several cluster sampling designs. Several bootstrap approaches are derived. A simulation study investigate the use of the derived approaches and a jackknife approach in the particular case of estimation of two diversity indices (Shannon index and Simpson index). A data analysis is presented (Chapter 4) when diversity is estimated for bird communities, using bootstrap methods.

CHAPTER 2. ESTIMATION OF SPECIES RICHNESS IN A BIOLOGICAL COMMUNITY

Introduction

The number of species - *species richness* - in a biological community is an important feature in the description of the community. It is considered as the simplest measure of diversity.

Among the problems inherent to the process of estimating the number of species in a biological community, a major one is the misrepresentation of rare species in a sample. Rare species occupy proportionally small areas within the community's boundaries. The definition of a sampling unit also constitutes a problem. In most cases, it is impossible to obtain random samples of individuals. Random selections of areas (or volumes) represent a practical alternative; that is, random selections of clusters of individuals are usually made. In the absence of some kind of natural cluster, dimensions and shapes for the sampling unit are arbitrarily defined; these sampling units are called *quadrats*. For example, Nilsson et al. (1988) estimated the number of species of land snails, in islands at Lake Mälaren, Sweden. A quadrat was defined as a 0.1 m^2 plot of land, from which all litter, the uppermost soil layer and all herbs were removed for subsequent analysis. Minshall et al. (1985) studied species richness of benthic invertebrate communities in streams by sampling rocks and

removing the invertebrates adhering from them. Although this provides a natural definition for a quadrat, a further restriction to sample only rocks with surface area of approximately 400 cm^2 was arbitrarily imposed.

The choice of a nonparametric estimator for the number of species seems reasonable, considering that the enormous variation among types of biological communities makes it difficult to formulate a parametric model suitable to every community. Resampling methods, such as the jackknife and the (nonparametric) bootstrap, are nonparametric techniques that only require a sample of quadrats selected through random sampling. The rest of this article deals with the estimation of number of species, when quadrat sampling is used.

Heltshe and Forrester (1983) developed a first order jackknife estimator for the number of species. They reached the conclusion that the jackknife estimator tends to underestimate the number of species when there are many rare species in the community. Special attention was given to the investigation of the effects of quadrat size in the estimation results. Simulations indicated that, for the same total sampled area, smaller quadrats gave better results. In this article an improved estimator for the variance of the jackknife estimator is proposed. Bootstrap techniques for estimating the number of species are also introduced. Simulation studies are used to compare the properties of confidence intervals and point estimators. An empirical Bayesian estimator, formulated by Mingoti (1989) is also investigated.

Estimators for the Number of Species Using Quadrat Sampling

In this section, applications of resampling methods to the estimation of the number of species - denoted S - in a biological community are addressed. Additional

information on jackknife and bootstrap methods may be found in Efron (1982), Miller (1974) and Efron and Gong (1983). Since only the first order jackknife is discussed, the term "jackknife" is used instead of "first order jackknife". All estimators considered here are based on a random sample of n quadrats. The notation $\hat{S}$ is used to denote the total number of species observed in a sample of quadrats. The last section presents a concise description of an empirical Bayesian estimator.

A jackknife method

Suppose that a quantity θ is estimated by $\hat{\theta}$, based on information from a random sample of size n. By deleting units one at a time, from the original sample (size n), n samples of size n-1 are created. Let $\hat{\theta}$ be the estimate derived from the sample of size n, and $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_n$ the estimates derived from the created samples. The jackknife estimator of θ based on $\hat{\theta}$, is then defined by

$$\hat{\theta}_{JK} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i, \tag{2.1}$$

where $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_i$, $i = 1, \ldots, n$. This method also provides an estimate for the variance of the jackknife estimator,

$$\widehat{Var}(\hat{\theta}_{JK}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_i - \hat{\theta}_{JK})^2, \tag{2.2}$$

which enables confidence intervals to be constructed for θ, in most cases, using Student's t distribution.

In the estimation of the number of species, S, Heltshe and Forrester (1983) applied the jackknife method to reduce the bias of $\hat{S}$ when a simple random sample of n quadrats is obtained. Closed form expressions for (2.1) and (2.2) were obtained,

$$\hat{S}_{JK} = \hat{S} + \frac{n-1}{n}R, \tag{2.3}$$

and

$$\widehat{Var}(\hat{S}_{JK}) = \frac{n-1}{n}\left\{\sum_{m=0}^{R} m^2 f_m - \frac{R^2}{n}\right\},$$

where R is the number of exclusive species (species found in only one quadrat) and f_m is the number of quadrats containing m exclusive species, $m = 1, \ldots, R$. An approximate $(1-\alpha)100\%$ confidence interval for S is given as

$$\left[\hat{S}_{JK} - t_{\frac{\alpha}{2},n-1}\sqrt{\widehat{Var}(\hat{S}_{JK})} \quad , \quad \hat{S}_{JK} + t_{\frac{\alpha}{2},n-1}\sqrt{\widehat{Var}(\hat{S}_{JK})}\right], \qquad (2.4)$$

where $t_{\alpha,n}$ is the $(1-\alpha)100^{th}$ percentile of Student's t distribution with n degrees of freedom. The approximation improves as n increases.

Bootstrap methods

Suppose that a random sample of size n is observed. A *bootstrap sample* is defined as a random sample of n units, sampled with replacement from the original sample of n units. A large quantity, B, of bootstrap samples are created. When a quantity θ is estimated by $\hat{\theta}$, from the original sample, each bootstrap sample also provides an estimate of θ. Confidence intervals for θ can be constructed in several ways, using bootstrap methods. An approximate $(1-\alpha)100\%$ confidence interval for θ is

$$\left[\hat{\theta}_{BT} - C_{\frac{\alpha}{2}}\hat{\sigma}_{BT} \quad , \quad \hat{\theta}_{BT} + C_{\frac{\alpha}{2}}\hat{\sigma}_{BT}\right],$$

where $\hat{\theta}_{BT}$ and $(\hat{\sigma}_{BT})^2$ are, respectively, the average and the sample variance of the B estimates from the bootstrap samples, and C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution. A simple procedure, called the "*percentile method* ", can be used to build a confidence interval for θ. The lower and upper

limits of the confidence interval are, respectively, the $(\frac{\alpha}{2})100^{th}$ percentile and the $(1-\frac{\alpha}{2})100^{th}$ percentile of the set of B estimates from the bootstrap samples.

In the application to the estimation of the number of species, based on a random sample of n quadrats, each bootstrap sample consists of a set of n quadrats selected with replacement from the original observed sample of n quadrats. The number of species in the original sample that are not present in the i^{th} bootstrap sample, m_i, is recorded for each bootstrap sample, i=1,$\cdots$,B. A $(1-\alpha)100\%$ confidence interval for the total number of species in the population, S, is

$$\left[\hat{S}+L \quad , \quad \hat{S}+U\right] , \tag{2.5}$$

where L and U are, respectively, the lower and upper $\alpha/2$ sample percentiles for $m_1,\ldots,m_B$.

Several bootstrap point estimates for the number of species can be defined:

$$E1 = \hat{S} \ + \ (\frac{U+L}{2}), \tag{2.6}$$

$$E2 = \hat{S} \ + \ (median\ of\ m_1,\ldots,m_B), \tag{2.7}$$

$$E3 = \hat{S} \ + \ (average\ of\ m_1,\ldots,m_B). \tag{2.8}$$

All three estimators are contained in the confidence interval defined in (2.5). If the histogram of $m_1,\ldots,m_B$ is symmetric about its center, the three estimators will be quite similar. If the histogram is skewed to the right, for example, E2 will tend to be smaller than both estimators E1 and E3. The estimator E3 coincides with an estimator developed by Smith and van Belle (1984). They used bootstrap methods to estimate the bias of $\hat{S}$ and make a suitable adjustment to $\hat{S}$. Closed form expressions were derived for the estimator and its variance. Consequently, the point

estimator $\hat{S} + E(m_i)$ can be computed without actually selecting bootstrap samples, but informative displays and the most reliable methods of constructing confidence intervals can not be completed without selecting the set of bootstrap samples.

Improved estimation for the variance of the jackknife estimator

This approach combines jackknife and bootstrap methods to get an improved estimator for the variance of $\hat{S}_{JK}$. Consider B bootstrap samples selected from the original sample of n quadrats. Let $\hat{S}_1, ..., \hat{S}_B$ be jackknife estimates for the number of species, S, obtained by applying (2.3) to each of the bootstrap samples. An estimator for the variance of $\hat{S}_{JK}$ is

$$\widehat{Var}_{Boot}(\hat{S}_{JK}) = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{S}_i - \bar{S})^2,$$

where $\bar{S} = \frac{1}{B} \sum_{i=1}^{B} \hat{S}_i$.

An approximate $(1-\alpha)100\%$ confidence interval for S is

$$\left[\hat{S}_{JK} - t_{\frac{\alpha}{2},n-1} \sqrt{\widehat{Var}_{Boot}(\hat{S}_{JK})} \quad , \quad \hat{S}_{JK} + t_{\frac{\alpha}{2},n-1} \sqrt{\widehat{Var}_{Boot}(\hat{S}_{JK})} \right], \tag{2.9}$$

where $t_{\alpha,n}$ is the $(1-\alpha)100$ percentile of Student's t distribution, with n degrees of freedom. The lower limit of (2.9) should be truncated at $\hat{S}$ if it is smaller than $\hat{S}$.

A Bayesian estimator

An empirical Bayesian estimator for the expected number of species was developed by Mingoti (1989), for a sample of n quadrats. Applications of this estimator to biological communities must consider the assumptions underlying the development of this estimator. The basis for the development of the estimator is modelling "the

probability that species s_i is observed in a typical quadrat of the region" as independent, identically distributed beta random variables, where $i = 1, ..., S$ and S is the total number of species in the community. It is also assumed that the number of quadrats in which species s_i is observed has a binomial distribution, conditional on the realized value of p_i, and these are independent across species. A particular model is also assumed for the distribution of the number of species that will be observed in a second independent sample of m quadrats but not observed in the first sample of n quadrats. The parameters α and β for the beta distribution function are unknown, and must be estimated, from information provided by the quadrat sample. There are no closed form expressions for the estimates of these parameters and numerical procedure to obtain values $\hat{\alpha}$ and $\hat{\beta}$ for those estimates. The final expression of an estimator for the expected number of species is

$$\hat{S}_{Bayes} = \hat{S} + \frac{n_1}{n\,\hat{\alpha}}(n + \hat{\beta} - 1)\left[1 - \frac{\Gamma(n + \hat{\alpha} + \hat{\beta})\,\Gamma(N + \hat{\beta})}{\Gamma(n + \hat{\beta})\,\Gamma(N + \hat{\alpha} + \hat{\beta})}\right], \tag{2.10}$$

where

n=sample size,

N=total number of quadrats,

$\hat{S}$=observed number of species, in the pool of n quadrats,

n_1=number of species found exclusively in one quadrat,

$\hat{\alpha}$=estimate of first parameter of the beta distribution function,

$\hat{\beta}$=estimate of second parameter of the beta distribution function,

$\Gamma(x)$=gamma function evaluated at x.

Compliance with the underlying assumptions for this estimator may be less than satisfactory for many biological communities. The underlying assumptions imply that individuals from every species tend to be evenly distributed across the study region, but many biological communities display spatial patchiness of species, with some species more abundant in particular areas than others. Different quadrats may contain patches for different species and other quadrats may be totally unsuitable for particular species. Deviations from the rather even distributions of species underlying the development of the Bayesian approach may seriously affect the applicability of this estimator. The Bayesian estimator is also a function of N (total number of quadrats). This might constitute a practical problem in applications to biological communities where the community boundaries are not well defined. In such cases N may not be known.

Illustration

Lloyd et al. (1968) sampled a Bornean rain forest in order to estimate a diversity index for some reptile and amphibian species. Their objective was not to estimate the number of species. In fact, it was known that 72 regular species reside in the particular region:

- *frogs* - 19 species,
- *lizards* - 18 species,
- *snakes* - 35 species.

From an area of 20 square miles, 402 quadrats, 25 × 25 ft areas, were randomly selected (therefore, 0.045% of total area was sampled). Since the information about

which species were found in each of the 402 quadrats was provided, estimates of the number of species can be compared with the true values.

Separate analyses for species of frogs, lizards and snakes, were performed. For the bootstrap estimation, one thousand bootstrap samples were created. Results are given for the estimation of the number of species of frogs, followed by species of lizards, and finally species of snakes. The histograms displayed in Figure 2.1 display some asymmetry. Confidence intervals defined in (2.5) might be more adequate to this situation rather than normal approximations.

From Table 2.1 it can be seen that there are three species of frogs found exclusively in a single quadrat. These three species appear in three different quadrats. Then, $\hat{S} = 18$, $R = 3$, $f_1 = 3$, $f_2 = 0$, $f_3 = 0$, and the jackknife estimate of S is

$$\hat{S}_{JK} = 18 + \left\{\frac{401}{402} \times 3\right\} = 21.0, \quad \text{with} \quad \widehat{Var}(\hat{S}_{JK}) = \frac{401}{402}\left\{3 - \frac{3^2}{402}\right\} = 2.97.$$

An approximate 90% confidence interval for S is

$$21 \pm 1.64\sqrt{2.97} = [18.2, 23.8] = [18, 24],$$

and, using $\widehat{Var}_{Boot}(\hat{S}_{JK}) = 5.11$, an approximate 90% confidence interval for S is

$$21 \pm 1.64\sqrt{5.11} = [17.3, 24.7] = [18, 25].$$

For the bootstrap samples summarized in Figure 2.1a,

$$\textit{average of estimates} = 1.8, \quad \textit{median of estimates} = 2,$$

$$5^{th}\ \textit{percentile} = 0 \quad \text{and} \quad 95^{th}\ \textit{percentile} = 4.$$

Therefore, the values of bootstrap point estimates for the number of species are E1 = 20, E2 = 20, E3 = 20, and a 90% bootstrap confidence interval for S, by the

Table 2.1: Species of frogs and number of quadrats in which they were found

Species	# of quadrats
Leptobrachium gracilis	102
Rana microdisca	19
Nesobia mjobergi	18
Kalophrynus pleurostigma	16
Ansonia longidigita	15
Bufo biporcatus	8
Ansonia albomaculata	7
Microhyla borneensis	7
Gastrophrynoides borneensis	5
Ichthyophis glutinosus	4
Leptobrachium gracilis	3
Kalophrynus intermedius	2
Microhyla annectens	2
Megophrys baluensis	2
Megophrys monticola	2
Calluela smithi	1
Micrixalus baluensis	1
Pelophryne brevipes	1

percentile method, is

$$[18+0, 18+4] = [18, 22].$$

Each interval estimate contains the true value S=19, in this case, but the bootstrap procedure offers more precision than the jackknife approach.

For the community of lizards, Table 2.2 shows that there is only one species found exclusively in one quadrat, among the 12 observed species. Then, $\hat{S} = 12$, $R = 1$, $f_1 = 1$, and the jackknife estimate for S is

$$\hat{S}_{JK} = 12 + \left\{\frac{401}{402} \times 1\right\} = 13.0 \text{ with } \widehat{Var}(\hat{S}_{JK}) = \frac{401}{402}\left\{1 - \frac{1^2}{402}\right\} = 1.00.$$

An approximate 90% confidence interval for S is

$$13 \pm 1.64\sqrt{1.00} = [11.4, 14.6] = [11, 15],$$

and, using $\widehat{Var}_{Boot}(\hat{S}_{JK}) = 3.43$, an approximate 90% confidence interval for S is

$$13 \pm 1.64\sqrt{3.43} = [10.0, 16.0] = [12, 16],$$

where 12 is used as lower limit since 12 species were observed in the sample. For the bootstrap samples summarized in Figure 2.1b,

$$\textit{average of estimates} = 1.0, \quad \textit{median of estimates} = 1,$$

$$5^{th} \textit{ percentile} = 0 \text{ and } 95^{th} \textit{ percentile} = 3.$$

Therefore, the values of bootstrap estimates, for the number of species are E1 = 14, E2 = 13, E3 = 13, and a 90% bootstrap confidence interval for S, by the percentile method, is

$$[12+0, 12+3] = [12, 15].$$

Table 2.2: Species of lizards and number of quadrats in which they were found

Species	# of quadrats
Aeluroscalabotes felinus	14
Phoxophrys nigrilabris	12
Mabuya rudis	7
Tropidophorus beccari	6
Cyrtodactylus pubisulcus	4
Dibamus novaeguineae	3
Gonyocephalus liogaster	3
Sphenomorphus cyanolaemus	3
Mabuya rubricollis	2
Mabuya multifasciatus	2
Sphenomorphus multiaquamatus	2
Tripidophorus micropus	1

The number of lizard species is underestimated by both the jackknife and bootstrap methods.

For the community of snakes, Table 2.3 shows three species found in a single quadrat. They appear in three different quadrats. Then, $\hat{S} = 8$, $R = 3, f_1 = 3$, $f_2 = 0$, $f_3 = 0$, and the jackknife estimate of S is

$$\hat{S}_{JK} = 8 + \left\{\frac{401}{402} \times 3\right\} = 11.0 \text{ with } \widehat{Var}(\hat{S}_{JK}) = \frac{401}{402}\left\{3 - \frac{3^2}{402}\right\} = 2.97.$$

An approximate 90% confidence interval for S is

$$11 \pm 1.64\sqrt{2.97} = [8.2, 13.8] = [8, 14],$$

and, using $\widehat{Var}_{Boot}(\hat{S}_{JK}) = 3.69$, an approximate 90% confidence interval for S is

$$11 \pm 1.64\sqrt{3.69} = [7.9, 14.2] = [8, 15].$$

For the bootstrap samples summarized in Figure 2.1c,

$$\textit{average of estimates} = 1.4, \quad \textit{median of estimates} = 1,$$

$$5^{th} \textit{ percentile} = 0 \text{ and } 95^{th} \textit{ percentile} = 3.$$

Therefore, the values of bootstrap point estimates for the number of species are E1 = 10, E2 = 9, E3 = 10, and a 90% bootstrap confidence interval for S, by the percentile method, is

$$[8 + 0, 8 + 3] = [8, 11].$$

In this case, both the jackknife and bootstrap estimates are well below the true value of S. This example is useful to illustrate three different situations which might be common in sampling animal communities:

Table 2.3: Species of snakes and number of quadrats in which they were found

Species	# of quadrats
Pseudorabdion collaris	5
Calamaria leucogaster	3
Maticora intestinalis	3
Pareas laevis	3
Natrix sarawakensis	2
Calamaria suluensis	1
Liopeltis baliodeirus	1
Liopeltis longicaudus	1

species of frogs: most of the species were observed;
every confidence interval considered contained the true number of species;

species of lizards: only 61% of the existing species were observed;
the confidence intervals are shifted slightly below the true number of species;

species of snakes: only 23% of the existing species were observed;
all confidence intervals lie far below the true number of species.

Estimates using the Bayesian method, described in Section (2.3), were also derived. In the case of frogs and snakes, where the number of species found in one exclusive quadrat was 3, the Bayesian estimate turned out to be the observed number of species increased by 22 and 28 species, respectively. As for the lizards, where only one species was located in an exclusive quadrat, the Bayesian estimator increased the observed

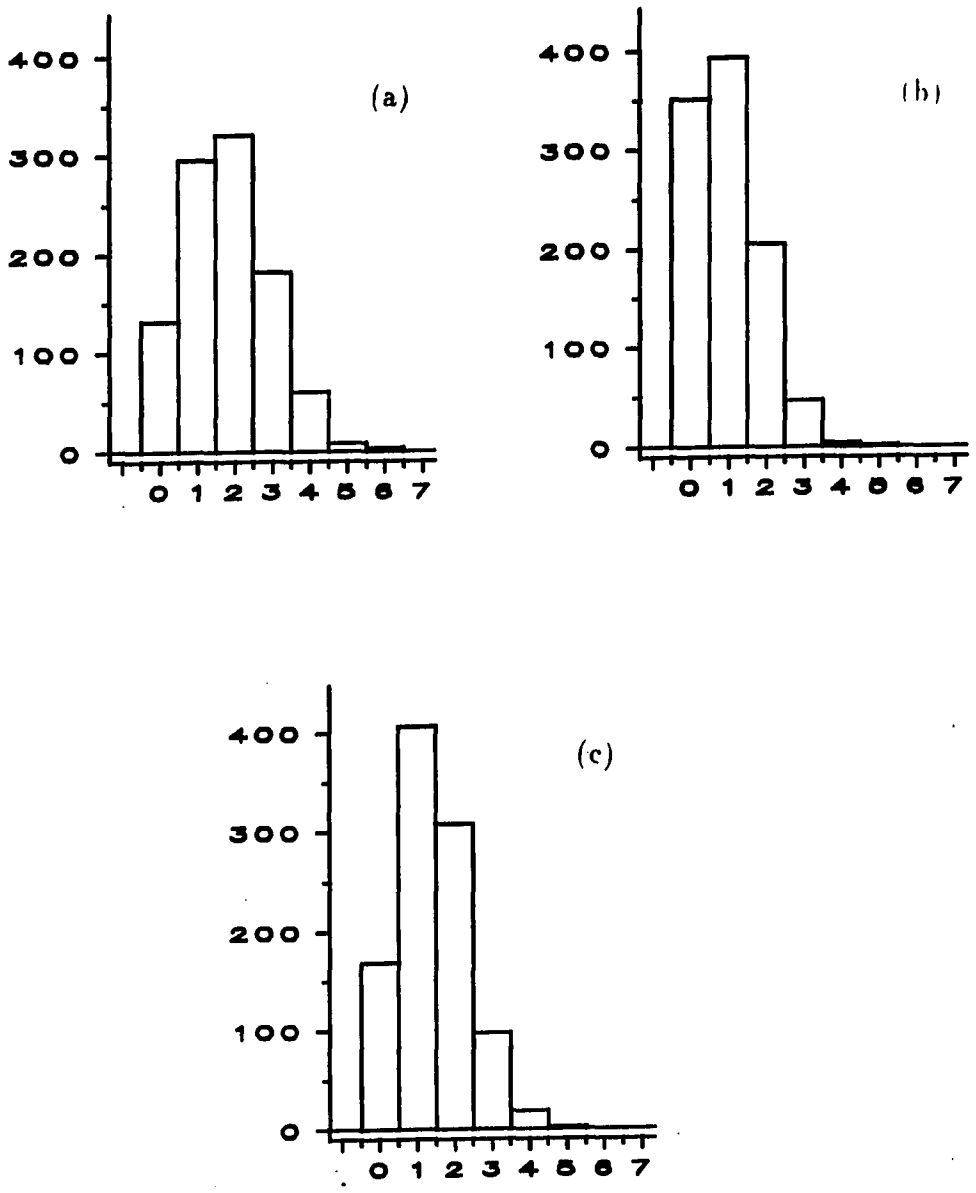


Figure 2.1: Histograms of number of species observed in the original sample but missing from bootstrap samples (1000 bootstrap samples) - (a) species of frogs, (b) species of lizards, (c) species of snakes

Table 2.4: Number of species of frogs, lizards and snakes (true and observed), jackknife estimate, $\hat{S}_{JK}$, bootstrap estimate, E1, and Bayesian estimate, $\hat{S}_{Bayes}$

	True	Observed	Jackknife	Bootstrap	Bayes
Frog	19	18	21	20	41
Lizard	18	12	13	14	13
Snake	35	8	11	10	36

number of species by less than one. Therefore, as shown in Table 2.4, the number of species of frogs was overestimated, the estimated number of species of snakes was very close to the true value and the number of species of lizards was underestimated.

The poor performance of the resampling methods for estimating the number of species of snakes can be at least partially attributed to problems with finding the snakes in the sample quadrats. Lloyd et al. (1968) note that: "our method of searching quadrats may somehow be specially inefficient for snakes. This could happen, for example, if certain snakes were to slip out of the quadrat unnoticed while we were just beginning to clear away the litter around the edge...". The resampling methods can be severely biased by systematically undercounting the number of species in some quadrats.

Simulation Design

Simulations, similar to those performed by Heltshe and Forrester (1983), were used to compare the properties of the various estimators for the number of species. Four communities, labelled A, B, C and D, each one with 25 species, were assembled

Table 2.5: Simulated communities

Community	A	B	C	D
Total number of individuals	3881	4308	4308	1038
Dispersion around parent (σ)	0.14	0.14	.02	.02
Offspring per parent	many	few	few	few

on unit squares. To generate patches of individuals, initially a parent was located randomly, and offspring were located around the parent; the distance of an offspring to a parent followed a normal distribution and the direction was determined from a uniform distribution on the angle from the east. The number of parents and offspring, and their dispersion were controlled by an initial choice of parameters. The same value was used for the dispersion parameter, σ, for all species within a community. The estimation of the number of species is least difficult for community A, designed to have many offspring per parent; communities B and C differ only in the dispersion of the offspring around parents and most of the individuals in these communities belong to one of five species. Therefore, B and C represent very uneven communities. Community D contains mostly scarce species with small dispersion among individuals in a group. Table 2.5 presents some features of the four communities.

One hundred samples were selected from each community, under each of nine quadrat sampling schemes, defined by three quadrat sizes and three totals for sampled area (the three quadrat sizes were defined by dividing the unit square into 196, 400 and 784 quadrats of same area). In earlier stages of this study, 200 bootstrap samples were created for each observed sample. However, recent applications reported in the literature have often used larger numbers of bootstrap samples. Therefore, all boot-

strap estimations were repeated with 1000 bootstrap samples. Results are presented only for the later case, 1000 bootstrap samples. Very similar results were obtained with 200 bootstrap samples. This suggests that there is no need to increase the number of bootstrap samples beyond 1000 (in this study). Point estimates and 90% confidence intervals for S were calculated, using the resampling procedures discussed previously. The Bayesian estimator (2.10) was evaluated for the first 20 samples of large quadrats selected from each community.

In the FORTRAN programs written to perform the simulations, all random number generation was done using IMSL subroutines.

Simulation Results

Results of the simulation study are displayed in 3×3 arrays with rows corresponding to quadrat size and columns corresponding to total area sampled. Each cell of the area contains the average point estimate for S, estimated mean square error of the estimate of S, average confidence interval length, and estimated coverage probabilities for confidence interval for both the bootstrap, E1, and the jackknife, $\hat{S}_{JK}$, estimators. This is described by the key at the bottom of each table. All confidence intervals have a nominal 90% coverage probability. Jackknife confidence intervals are evaluated according to (2.4) and the bootstrap confidence intervals are constructed from the simple percentile method that uses the lower and upper 5^{th} percentiles of 1000 bootstrap estimates as the interval limits.

Tables 2.6 through 2.9 summarize results for the bootstrap and jackknife estimators, E1 (2.6) and $\hat{S}_{JK}$ (2.3), from one hundred samples of quadrats, for each combination of quadrat size and area sampled and each community. Calculations

were performed for the three bootstrap estimators E1 (2.6), E2 (2.7), and E3 (2.8), but results for E2 and E3 are not reported since E1 consistently presented better results, in terms of smaller observed mean square error, although, for large total sampled areas, the confidence intervals and estimators did not differ much.

The bootstrap estimator tended to provide smaller estimates of S than the jackknife estimator of S. When many rare species are present in a community and relatively few species are observed, the bootstrap estimator severely underestimates the number of species. When more area is surveyed and more species are observed, the bootstrap estimator and related confidence intervals show dramatic improvement.

The jackknife estimator is a direct function of the number of rare species. For communities with many rare species, when only a small number are observed, the jackknife estimator tends to underestimate less than the bootstrap estimator. As the number of observed species increased, some jackknife estimates were larger than 25 (S), since there were still species observed exclusively in single quadrats. The coverage probabilities of jackknife confidence intervals show some improvement when the lower limit is changed to the closest integer below the computed value and the upper limit is changed to the closest integer above the computed value.

For a fixed amount of sampled area, changes in quadrat sizes had little effect on point estimates, although lengths of confidence intervals and coverage rates tend to increase as the quadrat size is increased. Increasing the amount of area surveyed has a more dramatic effect on improving the point estimators and confidence intervals. The bootstrap outperforms the jackknife procedure when larger percentages of area are surveyed.

The estimators for the variance of $\hat{S}_{JK}$, $\widehat{Var}(\hat{S}_{JK})$ and $\widehat{Var}_{Boot}(\hat{S}_{JK})$ are

conditioned on the observed sample. To assess the accuracy of these variance estimators, a new set of 1000 samples were selected from each community, for each combination of quadrat size and percent area surveyed. From each sample a jackknife point estimate, $\hat{S}_{JK}$, was calculated. The sample variance of the set of one thousand jackknife point estimates is an estimate for the unconditional variance of the jackknife estimator $\hat{S}_{JK}$. The averages of one hundred simulated values for both $\widehat{Var}(\hat{S}_{JK})$ and $\widehat{Var}_{Boot}(\hat{S}_{JK})$ were compared to the unconditional estimate. Results displayed in Tables 2.10, 2.11, 2.12 and 2.13 indicate that $\widehat{Var}_{Boot}(\hat{S}_{JK})$ tends to provide values closer to the unconditional variance of $\hat{S}_{JK}$. An explanation for this fact might be that bootstrap samples mimic a wider range of possible samples of quadrats, in comparison to the restricted samples created in the jackknife process.

Using $\widehat{Var}_{Boot}(\hat{S}_{JK})$, instead of $\widehat{Var}(\hat{S}_{JK})$, results in wider confidence intervals which seems to improve the percentage of coverage, as shown on Tables 2.14, 2.15, 2.16 and 2.17.

The Bayesian estimator, generally, has a larger variance than the resampling estimators. It tends to overestimate S even when larger areas are surveyed. For the smaller sample size, very frequently the number of species was greatly overestimated. In some samples of the smaller size, the Bayesian estimates were lower than respective estimates based on resampling methods, which also underestimate the number of species. The resampling estimators can be larger than S, but they are bounded above by 2S. There is no such upper bound for the Bayesian estimator.

In Figures 2.2, 2.3, 2.4 and 2.5 point estimates of S are plotted against the observed number of rare species in the sample, for the first 20 samples of large quadrats (selected from each community, for the three sample sizes). The Bayesian estimator

exhibits a strong linear relationship with the observed number of rare species in the sample. The resampling methods do not exhibit such a strong relationship. The Bayesian estimator was not defined for two samples of size 5 from community D, since there was no maximum for the function which determines its value.

Table 2.6: Observed results in the application of the jackknife estimator, $\hat{S}_{JK}$, and the bootstrap estimator, E1, in the estimation of S (S=25) - Community A

TOTAL SAMPLED AREA					
2.5%		5%		10%	
J	B	J	B	J	B
20 ▫		40 ▫		80 ▫	
24.7	22.2	25.4	24.0	25.5	24.9
10.8	11.1	4.2	2.2	2.7	1.3
7.4	4.8	5.3	3.5	3.5	2.2
73%	59%	78%	89%	80%	89%
10□		20□		40□	
24.7	21.9	25.5	24.0	25.8	25.0
9.8	11.4	5.9	2.9	2.9	1.1
7.9	5.3	5.7	3.8	4.1	2.5
77%	61%	76%	81%	81%	94%
5 □		10 □		20 □	
24.6	21.4	25.8	24.0	25.9	24.9
9.7	12.7	6.8	3.0	3.2	1.3
9.0	6.2	6.3	4.4	4.3	2.7
77%	65%	71%	86%	76%	94%

key

J	B	Jackknife	Bootstrap
sample size			
PE		average of point estimates	
MSE		observed mean square error	
CI Length		average of confidence interval lengths	
% COV		observed confidence interval coverage	
		(from 100 replications)	

Table 2.7: Observed results in the application of the jackknife estimator, $\hat{S}_{JK}$, and the bootstrap estimator, E1, in the estimation of S (S=25) - Community B

TOTAL SAMPLED AREA					
2.5%		5%		10%	
J	B	J	B	J	B
20 ▫		40 ▫		80 ▫	
21.9	18.8	25.1	22.8	26.1	24.7
17.4	39.4	7.1	7.6	7.0	2.1
8.0	4.8	7.0	4.5	5.2	3.4
65%	9%	78%	66%	69%	84%
10□		20□		40□	
22.8	19.2	25.1	22.5	26.0	24.7
16.5	35.4	10.9	10.0	5.1	2.0
8.4	5.4	7.5	4.8	5.5	3.7
65%	25%	72%	58%	81%	91%
5 □		10 □		20 □	
22.0	18.6	25.2	22.6	26.4	24.8
22.3	41.2	6.6	6.9	7.2	2.0
8.4	5.6	7.5	5.1	5.6	3.8
54%	22%	86%	76%	68%	91%

key

J	B	Jackknife	Bootstrap
sample size			
PE		average of point estimates	
MSE		observed mean square error	
CI Length		average of confidence interval lengths	
% COV		observed confidence interval coverage (from 100 replications)	

Table 2.8: Observed results in the application of the jackknife estimator, $\hat{S}_{JK}$, and the bootstrap estimator, E1, in the estimation of S (S=25) - Community C

TOTAL SAMPLED AREA					
2.5%		5%		10%	
J	B	J	B	J	B
20 ▫		40 ▫		80 ▫	
20.5	16.9	25.1	21.6	26.5	24.8
36.7	70.0	12.5	15.9	8.1	2.4
8.3	4.8	8.3	5.0	6.2	4.2
50%	5%	74%	50%	77%	90%
10□		20□		40□	
20.3	16.5	25.0	21.3	26.7	24.7
33.7	74.7	10.6	16.7	8.3	3.0
8.6	5.3	8.6	5.4	6.7	4.6
45%	4%	79%	49%	77%	89%
5 □		10 □		20 □	
19.1	15.4	24.4	20.7	27.4	24.8
48.3	87.9	11.4	20.7	13.7	3.4
8.7	5.7	8.9	5.7	7.3	5.0
31%	5%	78%	44%	67%	91%

key

J	B	Jackknife	Bootstrap
sample size			
PE		average of point estimates	
MSE		observed mean square error	
CI Length		average of confidence interval lengths	
% COV		observed confidence interval coverage	
		(from 100 replications)	

Table 2.9: Observed results in the application of the jackknife estimator, $\hat{S}_{JK}$, and the bootstrap estimator, E1, in the estimation of S (S=25) - Community D

TOTAL SAMPLED AREA					
2.5%		5%		10%	
J	B	J	B	J	B
20 ▫		40 ▫		80 ▫	
18.8	14.2	25.6	20.8	28.2	25.4
56.9	119.8	17.6	23.2	19.3	4.7
9.1	5.1	9.6	5.8	7.9	5.1
33%	2%	75%	42%	57%	86%
10□		20□		40□	
17.3	13.0	24.6	20.0	27.9	24.9
74.4	143.5	15.2	30.8	18.5	4.8
9.1	5.3	9.5	6.0	8.1	5.5
23%	0%	73%	37%	63%	90%
5 ☐		10 ☐		20 ☐	
15.8	12.1	23.3	18.5	28.0	24.3
104.6	167.0	19.3	45.2	23.2	4.4
9.2	5.6	10.2	6.5	8.6	5.9
21%	0%	72%	25%	67%	90%

key

J	B	Jackknife	Bootstrap
sample size			
PE		average of point estimates	
MSE		observed mean square error	
CI Length		average of confidence interval lengths	
% COV		observed confidence interval coverage	
		(from 100 replications)	

Table 2.10: Estimates for the variance of $\hat{S}_{JK}$ using three different estimators: (a) sample variance of 1000 point estimates $\hat{S}_{JK}$, (b) average of 100 bootstrap point estimates $\widehat{Var}_{Boot}(S_{JK})$ and (c) average of 100 jackknife point estimates $\widehat{Var}(S_{JK})$ - Community A

	Total Sampled Area		
	2.5%	5%	10%
▫	10.35	5.51	2.42
	7.21	3.96	2.42
	4.93	2.73	1.42
□	8.23	5.35	2.65
	9.32	4.63	2.43
	5.26	2.99	1.76
⊡	9.18	5.61	2.48
	13.88	6.15	2.72
	5.22	3.31	1.75

key	
quadrat size	sample variance of jackknife estimates average of bootstrap estimates average of jackknife estimates

Table 2.11: Estimates for the variance of $\hat{S}_{JK}$ using three different estimators: (a) sample variance of 1000 point estimates $\hat{S}_{JK}$, (b) average of 100 bootstrap point estimates $\widehat{Var}_{Boot}(S_{JK})$ and (c) average of 100 jackknife point estimates $\widehat{Var}(S_{JK})$ - Community B

	Total Sampled Area		
	2.5%	5%	10%
	12.61	8.20	4.50
▫	7.36	6.25	3.96
	5.66	4.51	2.77
	13.44	9.05	4.72
□	9.51	7.31	4.48
	5.60	5.03	2.91
	11.46	9.41	4.41
☐	11.97	8.52	4.60
	4.58	4.47	2.93

key	
quadrat size	sample variance of jackknife estimates
	average of bootstrap estimates
	average of jackknife estimates

Table 2.12: Estimates for the variance of $\hat{S}_{JK}$ using three different estimators: (a) sample variance of 1000 point estimates $\hat{S}_{JK}$, (b) average of 100 bootstrap point estimates $\widehat{Var}_{Boot}(S_{JK})$ and (c) average of 100 jackknife point estimates $\widehat{Var}(S_{JK})$ - Community C

	Total Sampled Area		
	2.5%	5%	10%
▫	14.46	12.00	6.24
	7.77	7.91	5.56
	6.01	6.26	3.76
□	15.11	12.05	6.58
	9.66	9.17	6.41
	5.77	6.47	4.16
☐	12.71	13.50	7.44
	12.40	11.18	7.79
	4.65	6.20	4.64

key	
quadrat size	sample variance of jackknife estimates average of bootstrap estimates average of jackknife estimates

Table 2.13: Estimates for the variance of $\hat{S}_{JK}$ using three different estimators: (a) sample variance of 1000 point estimates $\hat{S}_{JK}$, (b) average of 100 bootstrap point estimates $\widehat{Var}_{Boot}(S_{JK})$ and (c) average of 100 jackknife point estimates $\widehat{Var}(S_{JK})$ - Community D

	Total Sampled Area		
	2.5%	5%	10%
▫	19.28	15.30	8.50
	8.86	10.63	8.25
	7.22	8.46	5.79
□	17.56	16.80	8.97
	9.95	11.63	9.29
	6.56	7.92	6.08
□	17.40	16.92	9.54
	12.04	14.58	10.89
	5.29	8.22	6.45

key	
quadrat size	sample variance of jackknife estimates average of bootstrap estimates average of jackknife estimates

Table 2.14: Percentage of coverage of three types of confidence intervals for S (nominal coverage 90%): (a) jackknife method, (b) bootstrap method and (c) jackknife and bootstrap methods combined ($\widehat{S}_{JK}$ as estimator for S and $\widehat{Var}_{Boot}(S_{JK})$ as estimator for the variance of $\widehat{S}_{JK}$) - Community A

	Total Sampled Area		
	2.5%	5%	10%
▫	73 59 81	78 89 88	80 89 85
□	77 61 88	76 81 87	81 94 90
☐	77 65 95	71 86 94	76 94 92

key

quadrat size

jackknife method

bootstrap method

bootstrap and jackknife methods

Table 2.15: Percentage of coverage of three types of confidence intervals for S (nominal coverage 90%): (a) jackknife method, (b) bootstrap method and (c) jackknife and bootstrap methods combined ($\hat{S}_{JK}$ as estimator for S and $\widehat{Var}_{Boot}(S_{JK})$ as estimator for the variance of $\hat{S}_{JK}$) - Community B

	Total Sampled Area		
	2.5%	5%	10%
▫	65 9 69	78 66 84	69 84 75
□	65 25 77	72 58 86	81 91 88
☐	54 28 82	76 68 94	91 95 84

key

quadrat size

jackknife method

bootstrap method

bootstrap and jackknife methods

Table 2.16: Percentage of coverage of three types of confidence intervals for S (nominal coverage 90%): (a) jackknife method, (b) bootstrap method and (c) jackknife and bootstrap methods combined ($\widehat{S}_{JK}$ as estimator for S and $\widehat{Var}_{Boot}(S_{JK})$ as estimator for the variance of $\widehat{S}_{JK}$) - Community C

	Total Sampled Area		
	2.5%	5%	10%
▫	50 5 52	74 50 80	77 90 86
□	45 4 58	79 49 91	77 89 88
☐	31 5 58	78 44 87	67 91 81

key

quadrat size

jackknife method

bootstrap method

bootstrap and jackknife methods

Table 2.17: Percentage of coverage of three types of confidence intervals for S (nominal coverage 90%): (a) jackknife method, (b) bootstrap method and (c) jackknife and bootstrap methods combined ($\widehat{S}_{JK}$ as estimator for S and $\widehat{Var}_{Boot}(S_{JK})$ as estimator for the variance of $\widehat{S}_{JK}$) - Community D

	Total Sampled Area		
	2.5%	5%	10%
▫	33 2 40	75 42 83	57 86 68
□	23 0 31	73 37 85	63 90 78
☐	21 1 40	72 25 85	67 90 81

key

quadrat size

jackknife method

bootstrap method

bootstrap and jackknife methods

Conclusions

The simulation study and the real data problem showed that the estimators are less accurate when there are many rare species in a community. In those situations larger samples would be needed, but sampling large proportions of a biological community is usually not a practical solution, due to limitations imposed by numerous factors (such as destruction of site, time, cost, available personal, etc.). Conditional on additional knowledge that most species were observed, the variability among quadrats will determine an adequate confidence interval for the number of species when resampling estimators are used.

Literature Cited

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS 38, SIAM–NSF

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, 37, 36–48

Heltshe, J. F. and Forrester, N. E. (1983). Estimating Species Richness Using the Jackknife Procedure. Biometrics, 39, 1–11

Heltshe, J. F. and Forrester, N. E. (1985). Statistical Evaluation of the Jackknife Estimate of Diversity when Using Quadrat Samples. Ecology, 66, 107–111

Lloyd, M., Inger, R. F. and King, F. W. (1968). On the Diversity of Reptile and Amphibians Species in a Bornean Rain Forest. The American Naturalist, 102, 497–515

Miller, R. G. (1974). The Jackknife – a Review. Biometrica, 61, 1–15

Mingoti, S. A. (1989). Estimating the Total Number of Distinct Species When Quadrat Sampling or Sampling by Elements Is Used. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa

Minshall, G. W., Petersen, R. C. and Nimz, C. F. (1985). Species Richness in Streams of Different Size from the Same Drainage Basin. The American Naturalist, 125, 16–38

Morton, S. R. and Davidson, D. W. (1988) Comparative Structure of Harvester Ant Communities in Arid Australia and North America. Ecological Monographs, 58(1), 19–38

Nilsson, S. G., Bengtsson, J. and As, S. (1988). Habitat Diversity or Area Per Se? Species Richness of Woody Plants, Carabid Beetles and Land Snails on Islands. Journal of Animal Ecology, 57, 685–704

Pielou, E. C. (1975). Ecological Diversity. John Wiley, New York.

Smith, E. P. and van Belle, G. (1984). Nonparametric Estimation of Species Richness. Biometrics, 40, 119–129

CHAPTER 3. BOOTSTRAP ESTIMATION OF ECOLOGICAL DIVERSITY UNDER COMPLEX SAMPLING

Introduction

The estimation of a diversity index is usually based on a sample of quadrats (plots of land, volumes of water, etc.). The random selection of quadrats often involves multiple stages of selection or stratification. This study analyzes the estimation of a diversity index when several stages of clustering are employed. A diversity index is a function of the species proportions, and a proportion is a particular type of mean. Therefore, throughout this article, the more general problem of estimating a function of S population means, $f(\mu_1, \cdots, \mu_S)$ is considered. Estimates of f are usually obtained by evaluating f with the estimates of the means that are most suitable for the sampling design used in the study. While this can provide a reasonable point estimate $\hat{f} = f(\hat{\mu}_1, \cdots, \hat{\mu}_S)$, there usually is no closed form expression for the variance of $\hat{f}$.

For many functions, the Delta Method is useful in providing an approximation for the variance of $\hat{f}$,

$$Var(\hat{f}) \doteq \sum_{i=1}^{S} \sum_{j=1}^{S} \widehat{Cov}\left(\hat{\mu}_i, \hat{\mu}_j\right) f_i'(\hat{\mu}) f_j'(\hat{\mu}),$$

where $f_i'(\hat{\mu})$ is the first partial derivative of f with respect to the i^{th} coordinate,

evaluated at the point $\hat{\mu} = (\hat{\mu}_1, \cdots, \hat{\mu}_S)$. Alternatively, in the bootstrap estimation of f, individual means are estimated from a bootstrap sample, $\hat{\mu}_1^*, \cdots, \hat{\mu}_S^*$, and f is estimated by $f^* = f(\hat{\mu}_1^*, \cdots, \hat{\mu}_S^*)$. The expectation and variance of f^*, provide estimates of f and the variance of $\hat{f}$, respectively. The moments of f^* can be approximated by the sample moments of values of f^* obtained from a large number of bootstrap samples. Furthermore, the sample percentiles for a large number of bootstrap values f^* can be used to approximate percentiles of the distribution of $\hat{f}$. There are a number of ways of using the bootstrapped percentiles to construct a confidence interval for f.

Rao and Wu (1988) developed bootstrap estimators for functions of a population mean, f, for a stratified simple random sampling design and a two-stage cluster design. Modified versions of bootstrap methods were derived with the objective of obtaining consistent bootstrap estimates for the variance of $\hat{f}$. This results in using bootstrap sample sizes that are not necessarily the same as the original sample sizes for the stratified sampling. For the cluster sampling design, the number of clusters in a bootstrap sample is equal to the number of clusters selected at the first stage in the original sample. The number of selections from a resampled cluster is equal to the sample size of the original sample from the corresponding cluster, at the second stage. The resulting estimator for the two-stage cluster design is a function of the total number of elements in the population, and the total number of elements in the sampled clusters.

Cochran (1977) considered two basic estimators for a population mean under cluster sampling. One is unbiased, but it is a function of the total number of the elements in the population, which is usually unknown. The other estimator, the

sample average per element, does not require the knowledge of the number of elements in the population, although it is biased. The bootstrap estimators derived by Rao and Wu (1988) are based on unbiased estimators for the population means. This article presents bootstrap estimators based on biased estimators of population means (ratio-to-size estimators), that can be directly applied when the total number of elements in the population (community) is unknown. This is generally the case when ecological communities are sampled.

One-stage cluster sampling (corresponding to a random sample of quadrats) is considered first, and a basic bootstrap estimator based on the ratio-to-size estimator is defined. A modified estimator is derived following the techniques employed by Rao and Wu (1988). Finally, bias corrected versions for the bootstrap estimators are defined. A suggestion for stratified cluster sampling is presented. Estimators for diversity measures are compared through a simulation study. Aside from bootstrap estimators, a first order jackknife estimator considered by Heltshe and Forrester (1985) for the estimation of two diversity indices is also examined in this study. The results for one-stage cluster sampling are extended to cluster sampling with more than one stage, where at each stage clusters are subdivided into the same number of smaller clusters, and the clusters selected at the final stage of sampling are completely observed. This particular clustering corresponds to subdividing the study area into large quadrats of approximately same area, and subdividing them into the same amount of smaller quadrats of same area.

In the next Sections, the symbol * indicates estimates, expectations, variances, etc.,$\cdots$, related to the bootstrap sampling scheme. Superscripts labelling bootstrap samples are dropped, whenever possible, for simplicity. A subscript denoting variables

in a multivariate context is omitted in most basic definitions, since the definitions describe the univariate component.

Estimation of Population Means Using One-Stage Cluster Sampling

Consider the problem of estimating a population mean from a random sample of clusters where different clusters may contain different numbers of elements. The notation used follows the convention of capital letters for population values, lower case letters for random variables and sample design parameters, one bar to indicate average per cluster, and two bars to indicate average per element. Some commonly used symbols are:

N = number of clusters in the population,

M_i = number of elements in the i^{th} cluster, i=1,$\cdots$,N,

$\overline{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$ = mean number of elements per cluster,

Y_{ij} = j^{th} element, in the i^{th} cluster, j=1,$\cdots$,M_i, i=1,$\cdots$,N,

$Y_i = \sum_{j=1}^{M_i} Y_{ij}$ = total in the i^{th} cluster, i=1,$\cdots$,N,

$Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$ = population total,

$\overline{Y} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} \; / \; N$ = population mean per cluster,

$\overline{\overline{Y}} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} \; / \; \sum_{i=1}^{N} M_i$ = population mean per element,

n = number of clusters in the sample,

m_i = value of M_i for the i^{th} sampled cluster, i=1,..,n,

y_{ij} = j^{th} element in the i^{th} sampled cluster, j=1,$\cdots$,m_i, i=1,$\cdots$,n,

$y_i = \sum_{j=1}^{m_i} y_{ij}$ = total in the i^{th} sampled cluster, i=1,$\cdots$,n,

$\overline{y} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij} \; / \; n$ = sample average per cluster,

$\overline{\overline{y}} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij} \; / \; \sum_{i=1}^{n} m_i$ = sample average per element.

Cochran (1977) discussed two general estimators for a population total. One is the unbiased estimator

$$\hat{Y}_{\mathsf{U}} = N\overline{y}, \tag{3.1}$$

the sample average per cluster multiplied by the number of clusters in the population. The variance is

$$Var(\hat{Y}_{\mathsf{U}}) = N^2 \, [1 - (n/N)] \, \frac{S^2}{n},$$

where $S^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \; / \; (N-1)$. The other is a biased ratio-to-size estimator

$$\hat{Y}_{\mathsf{R}} = \left(\sum_{i=1}^{N} M_i \right) \overline{\overline{y}}, \tag{3.2}$$

the sample average per element multiplied by the total number of elements in the population. Although there is no closed form expression for the variance of (3.2), Cochran (1977) presented the approximation,

$$Var(\hat{Y}_{\mathsf{R}}) \doteq N^2 \, [1 - (n/N)] \, \frac{S_{\mathsf{R}}^2}{n}, \tag{3.3}$$

where $S_{\mathsf{R}}^2 = \sum_{i=1}^{N} (Y_i - M_i \overline{\overline{Y}})^2 \; / \; (N-1)$. Since $\overline{\overline{Y}} = Y \; / \; \sum_{i=1}^{N} M_i$, estimators for the population mean by element, $\overline{\overline{Y}}$, derived from (3.1) and (3.2), are

$$\hat{\overline{\overline{Y}}}_{\mathsf{U}} = \frac{1}{\sum_{i=1}^{N} M_i} \, (N \, \overline{y}) = \frac{\overline{y}}{\sum_{i=1}^{N} M_i \; / \; N},$$

and

$$\hat{\bar{\bar{Y}}}_R = \frac{1}{\sum_{i=1}^{N} M_i} \left(\sum_{i=1}^{N} M_i \, \bar{\bar{y}} \right) = \bar{\bar{y}} = \left(\frac{\bar{y}}{\sum_{i=1}^{n} m_i \, / \, n} \right),$$

respectively. $\hat{\bar{\bar{Y}}}_U$ is an unbiased estimator with variance

$$\begin{aligned} Var(\hat{\bar{\bar{Y}}}_U) &= \frac{1}{\left(\sum_{i=1}^{N} M_i/N\right)^2} Var(\hat{Y}_U) \\ &= \frac{1}{\left(\sum_{i=1}^{N} M_i/N\right)^2} [1 - (n/N)] \, \frac{S^2}{n}. \end{aligned}$$

$\hat{\bar{\bar{Y}}}_R$ is usually biased. Using (3.3) an approximation for its variance is

$$\begin{aligned} Var(\hat{\bar{\bar{Y}}}_R) &= \frac{1}{\left(\sum_{i=1}^{N} M_i \, / \, N\right)^2} Var(\hat{Y}_R) \\ &\doteq \frac{1}{\left(\sum_{i=1}^{N} M_i/N\right)^2} N^2 \, [1 - (n/N)] \, \frac{S_R^2}{n}. \end{aligned}$$

Cochran (1977) remarked that the precision of the unbiased estimator $\hat{Y}_U$ is relatively poor, especially when the cluster averages do not vary much, but the cluster sizes (M_i) have a large variation. This also applies to the unbiased estimator $\hat{\bar{\bar{Y}}}_U$. Another problem with the estimator $\hat{\bar{\bar{Y}}}_U$ is that $\sum_{i=1}^{N} M_i$ is frequently unknown. Cochran (1977) added that the biased estimator $\hat{Y}_R$ often has smaller variance than $\hat{Y}_U$ since it depends on variability among means by element. Expressions for $Var(\hat{\bar{\bar{Y}}}_U)$ and $Var(\hat{\bar{\bar{Y}}}_R)$ are functions of the total number of elements in the population.

An approximation for $Var(\hat{\bar{\bar{Y}}}_R)$ which does not depend on $\sum_{i=1}^{N} M_i$ can be obtained by interpreting the population mean by element as a ratio of two population means by cluster,

$$\bar{\bar{Y}} = \frac{(\sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij})/N}{(\sum_{i=1}^{N} M_i)/N}$$

$$= \overline{Y} \;/\; \overline{M}$$

$$= g(\overline{Y}, \overline{M}). \tag{3.4}$$

Under this approach, $g(\overline{Y}, \overline{M})$ is estimated by $g(\overline{y}, \overline{m}) = \overline{y} \;/\; \overline{m} = \hat{\overline{\overline{Y}}}_{\mathsf{R}}$, and the estimate of $Var(\hat{\overline{\overline{Y}}}_{\mathsf{R}})$ based on the Delta Method approximation for the variance of $g(\overline{y}, \overline{m})$ is based on estimates for the variances and covariance of $\overline{y}$ and $\overline{m}$,

$$\begin{aligned} \widehat{Var}(\overline{y}) &= [1-(n/N)]\,\frac{s_{\mathsf{y}}^2}{n}, \\ \widehat{Var}(\overline{m}) &= [1-(n/N)]\,\frac{s_{\mathsf{m}}^2}{n}, \\ \widehat{Cov}(\overline{y}, \overline{m}) &= [1-(n/N)]\,\frac{s_{\mathsf{ym}}}{n}. \end{aligned} \tag{3.5}$$

where $s_{\mathsf{y}}^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 \;/\; (n-1)$, $s_{\mathsf{m}}^2 = \sum_{i=1}^{n}(m_i - \overline{m})^2 \;/\; (n-1)$ and $s_{\mathsf{ym}} = \sum_{i=1}^{n}(y_i - \overline{y})(m_i - \overline{m}) \;/\; (n-1)$.

Bootstrap Estimators for a Function of Population Means Under One-Stage Cluster Sampling

The application of bootstrap methods to one stage cluster sampling creates a bootstrap sample by selecting n clusters, with replacement, from the original sample of n clusters. A large number, B, of bootstrap samples are selected. A bootstrap estimator for $f(\overline{\overline{Y}})$, based on $f(\overline{\overline{y}})$, can be computed from the B bootstrap samples in the following way. For each bootstrap sample calculate:

$$\overline{\overline{y}}^* = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m_i^*} y_{ij}^* \;/\; \sum_{i=1}^{n} m_i^* & \text{if } \sum_{i=1}^{n} m_i^* > 0 \\ 0 & \text{if } \sum_{i=1}^{n} m_i^* = 0, \end{cases} \tag{3.6}$$

and

$$f^* = f(\overline{\overline{y}}^*), \tag{3.7}$$

where m_i^* is the value of m_i and y_{ij}^* is the j^{th} element from the i^{th} cluster in the bootstrap sample, j=1,$\cdots$,m_i^* and i=1,$\cdots$,n. When $m_i^* = 0$, the corresponding cluster is empty. This can occur in ecological surveys, since the number of individuals in a quadrat is usually not known before the quadrat is sampled and inspected. A bootstrap estimator for $f(\overline{\overline{Y}})$ is the expected value of f^*. This is approximated by the sample average of f^* values for the B bootstrap samples,

$$E_*(f^*) \doteq \frac{1}{B}\sum_{j=1}^{B} f^{*j} = f^*_{\text{boot}}. \tag{3.8}$$

The bootstrap estimator for the variance of $f(\overline{\overline{y}})$ is the variance of f^* which is approximated by the sample variance of the set of f^* values from the B bootstrap samples,

$$Var_*(f^*) \doteq \frac{1}{B-1}\sum_{j=1}^{B}(f^{*j} - f^*_{\text{boot}})^2. \tag{3.9}$$

An approximate $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ is

$$\left[f^*_{\text{boot}} - C_{\frac{\alpha}{2}}(Var_*(f^*))^{1/2} \quad , \quad f^*_{\text{boot}} + C_{\frac{\alpha}{2}}(Var_*(f^*))^{1/2}\right],$$

where C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution. Alternatively, an $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ can be defined by the $(\frac{\alpha}{2})100^{th}$ and the $(1-\frac{\alpha}{2})100^{th}$ percentiles of the set of f^* values for the B bootstrap samples (percentile method). In this case, a point estimate for $f(\overline{\overline{Y}})$ can be defined as the median of the f^* values for the B bootstrap samples.

The above bootstrap estimator can also be developed with respect to the interpretation of $f(\overline{\overline{Y}})$ as $f(g(\overline{Y},\overline{M}))$, described in (3.4). For each bootstrap sample,

calculate:

$$\overline{y}^* = \begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{m_i^*} y_{ij}^* & \text{if } \sum_{i=1}^{n} m_i^* > 0, \\ 0 & \text{if } \sum_{i=1}^{n} m_i^* = 0, \end{cases}$$

$$\overline{m}^* = \sum_{i=1}^{n} m_i^*/n,$$

and

$$\begin{aligned} f_{\mathbf{g}}^* &= f[g(\overline{y}^*, \overline{m}^*)] \\ &= f^*. \end{aligned}$$

The evaluation of $g(.)$ uses the convention $0/0 = 0$. The final estimators for $f(\overline{\overline{Y}})$ and $Var[f(\overline{y})]$ are the same as those defined in (3.8) and (3.9), respectively. The variance of $\overline{y}^*$ under bootstrap sampling scheme is

$$\begin{aligned} Var_*(\overline{y}^*) &= Var_* \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i^*} y_{ij}^* \right) \\ &= \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2}{n} \\ &= \frac{n-1}{n} \frac{s_y^2}{n}, \end{aligned} \tag{3.10}$$

where s_y^2 is defined in (3.5). Therefore, $Var_*(\overline{y}^*)$ equals to $\widehat{Var}(\overline{y})$ (3.5) multiplied by $[(N-n)/N]^{-1}$ $[(n-1)/n]$. A modified estimator for a population mean by cluster with variance equal to $\widehat{Var}(\overline{y})$ can be obtained.

Consider a modified version of the bootstrap estimator where each bootstrap sample consists of n_b clusters selected with replacement from the original sample of n clusters. Recommendations for n_b are presented later. From each bootstrap sample

calculate:

$$\begin{aligned}
\overline{y}^* &= \frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{j=1}^{m_i^*} y_{ij}^*, \\
\overline{m}^* &= \frac{1}{n_b}\sum_{i=1}^{n_b} m_i^*, \\
\widetilde{\overline{y}}^* &= \overline{y} + \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\overline{y}^* - \overline{y}), \qquad (3.11) \\
\widetilde{\overline{m}}^* &= \overline{m} + \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\overline{m}^* - \overline{m}), \\
\widetilde{\overline{\overline{y}}}^* &= \widetilde{\overline{y}}^* \,/\, \widetilde{\overline{m}}^*, \qquad (3.12)
\end{aligned}$$

and

$$\tilde{f}^* = f(\widetilde{\overline{\overline{y}}}^*), \qquad (3.13)$$

where m_i^* is the number of elements and y_{ij}^* is the value for the j^{th} element in the i^{th} cluster of the bootstrap sample, j=1,$\cdots$,m_i^* and i=1,$\cdots$,n_b. The dependency on the value N might constitute a practical problem in applications to biological communities, since the exact value of N is rarely known. The expectation of $\widetilde{\overline{y}}^*$, under bootstrap sampling scheme, is

$$\begin{aligned}
E_*(\widetilde{\overline{y}}) &= E_*\left[\overline{y} + \left(\frac{N-n}{N}\frac{n_b}{n-1}\right)^{1/2}(\overline{y}^* - \overline{y})\right] \\
&= \overline{y} + \left(\frac{N-n}{N}\frac{n_b}{n-1}\right)^{1/2} E_*(\overline{y}^* - \overline{y}) \\
&= \overline{y}, \qquad (3.14)
\end{aligned}$$

where equation (3.14) follows (B.1) in Appendix B. The variance of $\widetilde{\overline{y}}^*$ under bootstrap sampling scheme is

$$Var_*(\widetilde{\overline{y}}) = Var_*\left[\overline{y} + \left(\frac{N-n}{N}\frac{n_b}{n-1}\right)^{1/2}(\overline{y}^* - \overline{y})\right]$$

$$= \left(\frac{N-n}{N}\frac{n_b}{n-1}\right) Var_*(\bar{y}^*)$$
$$= \widehat{Var}(\bar{y}),$$

which is given in (3.5). The modified bootstrap estimator for $f(\overline{\overline{Y}})$ is

$$\tilde{f}^*_{\text{boot}} = \frac{1}{B}\sum_{j=1}^{B} \tilde{f}^{*j}, \tag{3.15}$$

and the modified bootstrap estimator for the variance of $f(\overline{\overline{y}})$ is

$$Var_*(\tilde{f}^*) \doteq \frac{1}{B-1}\sum_{j=1}^{B}(\tilde{f}^{*j} - \tilde{f}^*_{\text{boot}})^2 . \tag{3.16}$$

Using the normal approximation, an approximate $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ is

$$\left[\tilde{f}^*_{\text{boot}} - C_{\frac{\alpha}{2}}(Var_*(\tilde{f}^*))^{1/2} \quad , \quad \tilde{f}^*_{\text{boot}} + C_{\frac{\alpha}{2}}(Var_*(\tilde{f}^*))^{1/2}\right],$$

where C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution, or, the percentile method might be used.

For a general function f (under some restrictions of continuity) of a multivariate mean of dimension S, (3.16) provides a consistent estimator for the variance of $f(\overline{\overline{y}})$. To show this, define

$$\bar{\mathbf{y}} = (\bar{y}_1, \cdots, \bar{y}_S),$$

$$\tilde{\mathbf{y}}^* = (\tilde{\bar{y}}^*_1, \cdots, \tilde{\bar{y}}^*_S),$$

$$\mathbf{d} = \tilde{\mathbf{y}}^* - \bar{\mathbf{y}} = (d_1, \cdots, d_S),$$

$$\hat{f} = f(\bar{\mathbf{y}}),$$

$$\tilde{f}^* = f(\tilde{\mathbf{y}}^*),$$

where the univariate component $\tilde{\bar{y}}_s^*$ is defined in (3.11), and $\bar{y}_s$ is a sample average by cluster, for s=1,$\cdots$,S. It is shown in (B.5) that $Var_*(d_s) = O_p(n^{-1})$. Therefore,

$$d_s = O_p(n^{-1/2}), \tag{3.17}$$

by Fuller (1976, page 186), for s=1,$\cdots$,S. By Fuller (1976, page 192), condition (3.17) implies that, for a function f with continuous third partial derivatives in a neighborhood of $\hat{\bar{\mathbf{y}}}$,

$$\tilde{f}^* = \hat{f} + \sum_{s=1}^{S} d_s f_s'(\bar{\mathbf{y}}) + \frac{1}{2}\sum_{s=1}^{S}\sum_{j=1}^{S} d_s d_j f_{sj}''(\bar{\mathbf{y}}) + O_p(n^{-3/2}), \tag{3.18}$$

where f_s' is the first partial derivative of f with respect to the s^{th} coordinate and f_{sj}'' is the second partial derivative of f with respect to the s^{th} and j^{th} coordinates. Using $E_*(d_s) = 0$, it follows that

$$E_*(\tilde{f}^* - \hat{f}) = \frac{1}{2}\sum_{s=1}^{S}\sum_{j=1}^{S} E_*(d_s d_j) f_{sj}''(\bar{\mathbf{y}}) + O_p(n^{-3/2}). \tag{3.19}$$

Then, since $E_*(d_s d_j) = O_p(n^{-1})$, as shown in (B.10),

$$\begin{aligned}\left[E_*(\tilde{f}^* - \hat{f})\right]^2 &= \left[\frac{1}{2}\sum_{s=1}^{S}\sum_{j=1}^{S} E_*(d_s d_j) f_{sj}''(\bar{\mathbf{y}}) + O_p(n^{-3/2})\right]^2 \\ &= \left[O_p(n^{-1}) + O_p(n^{-3/2})\right]^2 \\ &= O_p(n^{-2}).\end{aligned}$$

By (3.18),

$$E_*(\tilde{f}^* - \hat{f})^2 = E_*\left[\sum_{s=1}^{S} d_s f_s'(\bar{\mathbf{y}}) + \frac{1}{2}\sum_{s=1}^{S}\sum_{j=1}^{S} d_s d_j f_{ij}''(\bar{\mathbf{y}}) + O_p(n^{-3/2})\right]^2$$

$$= \sum_{s=1}^{S}\sum_{j=1}^{S} E_*(d_s d_j) f_s'(\overline{\mathbf{y}}) f_j'(\overline{\mathbf{y}}) + O_p(n^{-2}) \tag{3.20}$$

$$= \sum_{s=1}^{S}\sum_{j=1}^{S} \widehat{Cov}\left(\overline{y}_s, \overline{y}_j\right) f_s'(\overline{\mathbf{y}}) f_j'(\overline{\mathbf{y}}) + O_p(n^{-2}), \tag{3.21}$$

where (3.20) follows from $E_*[d_s d_j d_k] = O_p(n^{-2})$ and $E_*[d_s d_j d_k d_i] = O_p(n^{-3})$ shown in (B.12) and (B.13), respectively. Equation (3.21) follows from $E_*[d_s d_j] = \widehat{Cov}\left[\overline{y}_s\ ,\ \overline{y}_j\right]$, shown in (B.9). Consequently, the variance of $\tilde{f}^*$ under bootstrap sampling scheme is

$$\begin{aligned} Var_*(\tilde{f}^*) &= E_*(\tilde{f}^* - \hat{f})^2 + \left[E_*(\tilde{f}^* - \hat{f})\right]^2 \\ &= \sum_{s=1}^{S}\sum_{j=1}^{S} \widehat{Cov}\left(\overline{y}_s, \overline{y}_j\right) f_s'(\overline{\mathbf{y}}) f_j'(\overline{\mathbf{y}}) + O_p(n^{-2}). \end{aligned}$$

The first term of $Var_*(\tilde{f}^*)$ is the asymptotic expression for the variance of $\hat{f}$ provided by the Delta Method. Therefore, $\tilde{f}^*$ is a consistent estimator for $f(\overline{\overline{y}})$.

Bias corrected bootstrap estimators

The estimator $f(\overline{\overline{y}})$ is not necessarily an unbiased estimator of $f(\overline{\overline{Y}})$. A correction is introduced by subtracting a bootstrap estimate of bias from $f(\overline{\overline{y}})$.

The bias of $f(\overline{\overline{y}})$ can be estimated, from each bootstrap sample, by

$$b^* = f^* - f(\overline{\overline{y}}),$$

where f^* is defined in (3.7), and a bias corrected estimate for $f(\overline{\overline{Y}})$, from each bootstrap sample is

$$f(\overline{\overline{y}}) - b^*. \tag{3.22}$$

Then, for a set of B bootstrap samples, a bias corrected bootstrap estimator for $f(\overline{\overline{Y}})$, based on $f(\overline{\overline{y}})$, is

$$f^*_{bc} = f(\overline{\overline{y}}) - \frac{1}{B}\sum_{j=1}^{B} b^{*j}. \tag{3.23}$$

A bootstrap estimator for the variance of $f(\overline{\overline{y}})$ is

$$Var_*(f(\overline{\overline{y}}) - b^*) = Var_*(f^*),$$

which is defined in (3.9). An approximate $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ is

$$\left[f^*_{bc} - C_{\frac{\alpha}{2}}(Var_*(f^*))^{1/2} \quad , \quad f^*_{bc} + C_{\frac{\alpha}{2}}(Var_*(f^*))^{1/2}\right],$$

where C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution. Alternatively, an $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ can be defined by the $(\frac{\alpha}{2})100^{th}$ and the $(1-\frac{\alpha}{2})100^{th}$ percentiles of the set of $f(\overline{\overline{y}}) - b^*$ values for the B bootstrap samples.

Similarly, a bias corrected version for the modified bootstrap estimator for $f(\overline{\overline{Y}})$ is obtained by calculating, from each bootstrap sample, an estimate for the bias of $f(\overline{\overline{y}})$,

$$\tilde{b}^* = \tilde{f}^* - f(\overline{\overline{y}}),$$

where $\tilde{f}^*$ is defined in (3.13), and a bias corrected estimate for $f(\overline{\overline{Y}})$, from each bootstrap sample is

$$f(\overline{\overline{y}}) - \tilde{b}^*. \tag{3.24}$$

Then, for a set of B bootstrap samples, a bias corrected modified bootstrap estimator

for $f(\overline{\overline{Y}})$, based on $f(\overline{\overline{y}})$, is

$$\tilde{f}^*_{bc} = f(\overline{\overline{y}}) - \frac{1}{B}\sum_{j=1}^{B} \tilde{b}^{*j}. \tag{3.25}$$

A bootstrap estimator for the variance of $f(\overline{\overline{y}})$ is

$$Var_*(f(\overline{\overline{y}}) - \tilde{b}^*) \;=\; Var_*(\tilde{f}^*),$$

which is defined in (3.16). Thus, $\tilde{f}^*_{bc}$ is a consistent estimator of $f(\overline{\overline{y}})$. An approximate $(1-\alpha)100\%$ confidence interval for $f(\overline{\overline{Y}})$ is

$$\left[\tilde{f}^*_{bc} - C_{\frac{\alpha}{2}}(Var_*(\tilde{f}^*))^{1/2} \;\;,\;\; \tilde{f}^*_{bc} + C_{\frac{\alpha}{2}}(Var_*(\tilde{f}^*))^{1/2}\right],$$

where C_α is the $(1-\alpha)100^{th}$ percentile from the standard normal distribution. Alternatively, the $(\frac{\alpha}{2})100^{th}$ and the $(1-\frac{\alpha}{2})100^{th}$ percentiles of the set of $f(\overline{\overline{y}}) - \tilde{b}^*$ values for the B bootstrap samples can be used.

Bootstrap Estimators for a Function of Proportions under One Stage Cluster Sampling Design

Consider a population where the elements are classified into S categories. In the ecological context, a population is a *community* and the categories are *species*. Using the notation previously described (with a third subscript for the variable identification), let

$Y_{ijs} = 1$ if the j^{th} element from the i^{th} cluster belong to category s,
$= 0$ otherwise,

for j=1,$\cdots$,M_i, i=1,$\cdots$,N and s=1,$\cdots$,S. For this particular assignment of values, $\overline{\overline{Y}}_s$ is the proportion of elements in the population that belongs to category s, $\overline{\overline{y}}_s$ is the

sample proportion of elements belonging to category s, etc.,.... These would correspond to the species abundances in a community. The bootstrap estimators described in the former section can be applied to the estimation of a function of proportions. The original bootstrap estimator for a proportion, $\overline{\overline{y}}^*$ (3.6), is always a value in the interval [0,1]. However, the modified estimator, $\tilde{\overline{\overline{y}}}^*$ (3.12), can produce negative values. The size n_b of a bootstrap sample can be defined in order to guarantee that the modified estimator is a value in the interval [0,1].

Result 1 *The estimator* $\tilde{\overline{\overline{y}}}^*_s$ *in (3.12) is a value in the interval [0,1] if*

$$n_b \leq [(n-1)N/(N-n)],$$

where $[x]$ *is the larger integer contained in* x.

Proof: Let

$$K = \left\{ \frac{N-n}{N} \frac{n_b}{n-1} \right\}^{1/2}.$$

The condition

$$n_b \leq [(n-1)N/(N-n)]$$

is equivalent to $K \leq 1$. When $K \leq 1$,

$$\begin{aligned} \tilde{\overline{\overline{y}}}^*_s &= \frac{\overline{y}_s + K(\overline{y}^*_s - \overline{y}_s)}{\overline{m} + K(\overline{m}^* - \overline{m})} \\ &= \frac{\overline{y}_s(1-K) + K\overline{y}^*_s}{\overline{m}(1-K) + K\overline{m}^*} \\ &\geq 0, \end{aligned} \tag{3.26}$$

since $\overline{y}_s$, $\overline{y}^*_s$, $\overline{m}$ and $\overline{m}^*$ are non-negative. Summing across categories yields

$$\sum_{s=1}^{S} \tilde{\overline{\overline{y}}}^*_s = \sum_{s=1}^{S} \frac{(1-K)\overline{y}_s + K\overline{y}^*_s}{(1-K)\overline{m} + K\overline{m}^*}$$

$$= \frac{\sum_{s=1}^{S}\left\{(1-K)\overline{y}_s + K\overline{y}_s^*\right\}}{(1-K)\overline{m} + K\overline{m}^*}$$
$$= \frac{(1-K)\sum_{s=1}^{S}\overline{y}_s + K\sum_{s=1}^{S}\overline{y}_s^*}{(1-K)\overline{m} + K\overline{m}^*}. \qquad (3.27)$$

Applying the following results,

$$\begin{aligned}\sum_{s=1}^{S}\overline{y}_s &= \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ijs} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{s=1}^{S} y_{ijs} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} 1 \\ &= \frac{1}{n}\sum_{i=1}^{n} m_i \\ &= \overline{m}\end{aligned}$$

and,

$$\begin{aligned}\sum_{s=1}^{S}\overline{y}_s^* &= \frac{1}{n_b}\sum_{s=1}^{S}\sum_{i=1}^{n_b}\sum_{j=1}^{m_i^*} y_{ijs}^* \\ &= \overline{m}^*,\end{aligned}$$

to equation (3.27), yields

$$\sum_{s=1}^{S}\tilde{\overline{y}}_s^* = 1. \qquad (3.28)$$

By (3.27) and (3.28), $0 \le \tilde{\overline{y}}_s^* \le 1$, for s=1,$\cdots$,S. □

The above result provides an upper bound for the size of the bootstrap sample for the modified estimator of $f(\overline{\overline{Y}})$, which guarantees estimates for the proportions in the interval [0,1]. Bootstrap samples of greater size than the upper bound might

yield negative values for the estimate of proportions, as verified in a simulation study (described in the following section). The larger the sample fraction (n/N), the higher the upper bound for n_b. The smaller upper bound is n-1. Defining n_b as close as possible to the original sample size,

$$n_b = min\left\{n\ ,\ \frac{N}{N-n}(n-1)\right\},$$

results in using either as n or n-1 for n_b. If n_b is defined as the upper bound, the modified bootstrap estimator is defined as the original bootstrap estimator, (3.6), except that the bootstrap sample size is not necessarily equal to n.

Bootstrap Estimators for Diversity Indices under One-Stage Cluster Sampling Design

Simulation design

Simulations were performed to investigate bootstrap estimation of diversity indices. A community containing ten species was generated on a unit square, with random numbers and random locations of patches of individuals, for each of the species. The procedure used was similar to the one performed by Heltshe and Forrester (1983), where "parents" are located at random, and "offspring" are located around each parent. The numbers of parents, the number of offspring, and the location of offspring were determined as in Heltshe and Forrester (1983). The locations of parents were made differently, in order to introduce dependency among some species. The locations of parents for species 1 (the species with the largest number of parents) were randomly assigned according to the uniform distribution on the unit square. For species 2 and 3, the locations of parents were determined by adding random variables

distributed as a uniform [-0.1,0.1] to the coordinates of parents of species 1, until the specific number of parents of species 2 and 3 were reached. Similarly, species 4, 5 and 6 were generated. The locations of parents, for species 7 were determined according to the uniform distribution on the unit square. To locate a parent of species 8, a random variable following the uniform distribution on the unit square was initially observed; it was used if that point was not located within circles of radius 0.1 centered at the locations of parents of the first species, otherwise, another random variable was generated. Similarly, species 9 and 10 were generated. Therefore, two sets of species display positive association, {1,2,3} and {4,5,6}, and two sets of species display negative association, {7,8} and {9,10}. Three values were given for the dispersion parameter: for species 1, $\sigma = 0.05$, for species 3, $\sigma = 0.01$ and, for all other species, $\sigma = 0.02$. This community is characterized by dominance of species 1,2 and 3, and by small dispersion among individuals within a species.

For the sampling process, the unit square was subdivided into square quadrats of equal area. Three quadrat sizes were used, corresponding to grids of 14×14, 20×20 and 28×28 quadrats. The sample sizes considered correspond to sampling 2.5% , 5% and 10% of the unit square. For each one of the nine sampling schemes, defined by quadrat size and percentage of area sampled, one hundred independent samples of quadrats were selected from the community.

Point estimates and 90% confidence intervals for the Shannon index,

$$H' = -\sum_{i=1}^{S} \pi_i log(\pi_i),$$

and the Simpson index,

$$D = 1 - \sum_{i=1}^{S} \pi_i^2,$$

were calculated, using the four bootstrap methods described previously: original, modified, original bias corrected and modified bias corrected. In the construction of confidence intervals for the indices, normal approximation and percentile method were used. Therefore, eight types of confidence intervals were calculated. When normal approximation was used, the point estimates were

- f^*_{boot}, (3.8),
- $\tilde{f}^*_{\text{boot}}$, (3.15),
- f^*_{bc}, (3.23), and
- $\tilde{f}^*_{\text{bc}}$, (3.25),

respectively for the four methods cited above. The medians of sets of 1000 estimates

- f^*, (3.7),
- $\tilde{f}^*$, (3.13),
- $f(\overline{\overline{y}}) - b^*$, (3.22), and
- $f(\overline{\overline{y}}) - \tilde{b}^*$, (3.24),

from 1000 bootstrap samples, were used as point estimates, when using the percentile method, respectively, for the four methods cited above.

For the modified bootstrap estimators, bootstrap samples of size n or n-1 were created, depending on the upper bound of Result 1. A jackknife estimator, described in Heltshe and Forrester (1985) was also applied.

Simulation results

Tables 3.1 through 3.3 summarize the main features of the bootstrap estimation of the Simpson index, for the three sizes of selected quadrats, and Tables 3.4 through 3.6 for the Shannon index, for the bootstrap estimation. Tables 3.7 through 3.9 summarize the results of the jackknife estimation for both indices. The main evidence shown in those tables is that the correction for bias improved the bootstrap estimation, even for large samples, and that the results were very similar for the bias corrected bootstrap and the jackknife estimators. Analyzing the data for the bias corrected bootstrap, the modified version exhibited a moderate improvement only for samples of size 5 (large quadrats), for both indices. There was not a clear difference with the other sample sizes, regardless of quadrat size, as to a better performance by either the original or the modified version of the bootstrap. The jackknife estimates for the Shannon index are closer to the index value (2.07), but the confidence intervals are wider than all bootstrap versions. The jackknife point estimates for the Simpson index are very precise, as well as the bootstrap estimation results, as more quadrats are sampled.

The use of sample size n for bootstrap sample size, for the modified method, yielded many negative values for estimates of proportions, when n was higher than the upper bound given in Result 1. Changing the bootstrap sample size to $n-1$, gave proper values for the estimates.

Table 3.1: Observed results in the bootstrap estimation of the Simpson index (0.85) - large quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	5 □		10 □		20 □	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
P	.744	.819	.792	.839	.819	.844
	.015	.005	.005	.002	.002	.000
	.221	.221	.133	.133	.087	.087
	4%	93%	22%	91%	48%	90%
Pm	.715	.847	.770	.861	.819	.843
	.023	.004	.008	.001	.002	.000
	.279	.279	.154	.154	.084	.084
	3%	87%	3%	83%	49%	90%
N	.725	.837	.784	.846	.815	.848
	.020	.004	.006	.002	.002	.001
	.240	.239	.139	.139	.089	.089
	47%	92%	56%	95%	68%	94%
Nm	.688	.874	.760	.870	.816	.847
	.032	.004	.01	.002	.002	.001
	.332	.332	.164	.164	.086	.086
	50%	97%	37%	94%	67%	94%

key

B	B_{bc}	
PE		average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

B = Bootstrap B_{bc} = Bias-Corrected Bootstrap

P=percentile P_m=modified percentile

N=normal N_m=modified normal

Table 3.2: Observed results in the bootstrap estimation of the Simpson index (0.85) - medium quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	10□		20□		40□	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
	.748	.827	.780	.844	.824	.848
P	.014	.005	.004	.001	.001	.000
	.204	.204	.122	.122	.080	.080
	6 %	89%	31%	91%	61%	93%
	.712	.863	.800	.844	.825	.847
Pm	.024	.003	.004	.001	.001	.000
	.277	.277	.122	.122	.076	.076
	2%	83 %	31 %	91%	61%	92%
	.734	.841	.793	.851	.821	.851
N	.017	.004	.004	.001	.001	.001
	.217	.217	.127	.127	.082	.082
	48%	95%	64%	92 %	81%	94%
	.690	.885	.793	.851	.822	.850
Nm	.033	.004	.004	.001	.001	.001
	.313	.313	.127	.127	.078	.078
	41%	99 %	64%	92%	80%	92%

key

B	B_{bc}	
PE		average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

B Bootstrap B_{bc} Bias-Corrected Bootstrap

P=percentile P_m=modified percentile

N=normal N_m=modified normal

Table 3.3: Observed results in the bootstrap estimation of the Simpson index (0.85) - small quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	20 ▫		40 ▫		80 ▫	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
	.763	.826	.807	.843	.834	.853
P	.010	.004	.003	.001	.001	.001
	.172	.172	.108	.108	.063	.063
	9 %	93 %	36%	88%	75%	86%
	.705	.883	.808	.843	.835	.852
Pm	.026	.003	.003	.001	.001	.001
	.272	.272	.107	.107	.060	.060
	1 %	82 %	36 %	87%	74%	86%
	.753	.836	.802	.849	.831	.856
N	.012	.003	.003	.001	.001	.001
	.180	.180	.111	.111	.064	.064
	36%	94%	65%	93%	78%	89%
	.687	.902	.802	.848	.832	.855
Nm	.032	.05	.003	.001	.001	.001
	.299	.299	.110	.110	.060	.060
	34%	95%	65%	93%	78%	89%

key

B	B_{bc}	Bootstrap Bias-Corrected Bootstrap
PE		average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

P=percentile P_m=modified percentile

N=normal N_m=modified normal

Table 3.4: Observed results in the bootstrap estimation of the Shannon index (2.07) - large quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	5 ☐		10 ☐		20 ☐	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
	1.55	1.87	1.77	2.00	1.90	2.04
P	.30	.10	.11	.02	.03	.01
	.65	.65	.46	.46	.33	.33
	3%	79%	12%	88%	36%	92%
	1.53	1.89	1.80	1.97	1.91	2.04
Pm	.33	.09	.09	.03	.03	.01
	.71	.71	.48	.48	.31	.31
	3%	83%	23%	91%	36%	88%
	1.51	1.91	1.75	2.02	1.90	2.05
N	.35	.08	.12	.02	.04	.01
	.68	.68	.47	.47	.33	.33
	8%	77%	24%	89%	42%	89%
	1.48	1.94	1.78	1.99	1.90	2.04
Nm	.39	.08	.10	.03	.04	.01
	.76	.76	.49	.49	.32	.32
	9%	81%	32%	88%	41%	87%

key

B	B_{bc}	
PE		average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

B = Bootstrap; B_{bc} = Bias-Corrected Bootstrap

P=percentile; P_m=percentile-modified

N=normal; N_m=normal-modified

Table 3.5: Observed results in the bootstrap estimation of the Shannon index (2.07) - medium quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	10□		20□		40□	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
	1.58	1.91	1.81	2.03	1.93	2.06
P	.27	.08	.08	.02	.03	.01
	.60	.60	.43	.43	.30	.30
	3 %	78%	18%	84%	44%	89%
	1.60	1.89	1.81	2.03	1.94	2.06
Pm	.26	.09	.08	.02	.02	.01
	.65	.65	.43	.43	.29	.29
	5%	79%	18%	84%	44%	87%
	1.55	1.94	1.79	2.05	1.93	2.07
N	.30	.08	.09	.02	.03	.01
	.61	.61	.43	.43	.31	.31
	4 %	78 %	28 %	85%	54%	90%
	1.57	1.92	1.79	2.05	1.93	2.06
Nm	.29	.08	.09	.02	.03	.01
	.67	.67	.43	.43	.29	.29
	17%	78%	28%	85%	52%	90%

key

B	B_{bc}	
PE		average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

B Bootstrap B_{bc} Bias-Corrected Bootstrap

P=percentile P_m=percentile-modified

N=normal N_m=normal-modified

Table 3.6: Observed results in the bootstrap estimation of the Shannon index (2.07) - small quadrats

	TOTAL SAMPLED AREA					
	2.5%		5%		10%	
sample size	20 ▫		40 ▫		80 ▫	
	B	B_{bc}	B	B_{bc}	B	B_{bc}
	1.65	1.94	1.85	2.03	1.97	2.08
P	.20	.05	.06	.02	.02	.01
	.53	.53	.39	.39	.25	.25
	3%	78%	24%	87%	65%	82 %
	1.71	1.88	1.85	2.03	1.98	2.07
Pm	.16	.07	.06	.02	.01	.01
	.60	.60	.38	.38	.24	.24
	24%	79 %	25%	85 %	64%	78 %
	1.63	1.96	1.83	2.05	1.97	2.08
N	.22	.05	.07	.02	.02	.01
	.54	.54	.39	.39	.26	.26
	7%	77 %	35 %	88 %	66%	83 %
	1.69	1.90	1.84	2.04	1.97	2.08
Nm	.17	.06	.06	.02	.02	.01
	.61	.61	.38	.38	.24	.24
	32%	80%	35%	88 %	66%	82 %

key

B	B_{bc}	
PE		Bootstrap average of point estimates
MSE		observed mean square error
CI Length		average of confidence interval lengths
% COV		observed confidence interval coverage (from 100 replications)

P=percentile P_m=percentile-modified

N=normal N_m=normal-modified

Table 3.7: Observed results in the jackknife estimation of the Simpson index (0.85) and Shannon index (2.07) - large quadrats

	TOTAL SAMPLED AREA		
	2.5%	5%	10%
sample size	5 ☐	10 ☐	20 ☐
SIMPSON	.85 .005 .258 86%	.85 .002 .142 88%	.85 .001 .088 89%
SHANNON	2.04 .081 .885 81%	2.07 .023 .558 91%	2.06 .010 .368 92%

key

PE	average of point estimates
MSE	observed mean square error
CI Length	average of confidence interval lengths
% COV	observed confidence interval coverage

(from 100 replications)

Table 3.8: Observed results in the jackknife estimation of the Simpson index (0.85) and Shannon index (2.07) - medium quadrats

	TOTAL SAMPLED AREA		
	2.5%	5%	10%
sample size	10□	20□	40□
SIMPSON	.858	.855	.853
	.004	.001	.001
	.227	.123	.080
	86%	89%	94%
SHANNON	2.05	2.08	2.08
	.082	.025	.008
	.787	.501	.366
	80%	87%	92%

key	
PE	average of point estimates
MSE	observed mean square error
CI Length	average of confidence interval lengths
% COV	observed confidence interval coverage
	(from 100 replications)

Table 3.9: Observed results in the jackknife estimation of the Simpson index (0.85) and Shannon index (2.07) - small quadrats

	TOTAL SAMPLED AREA		
	2.5%	5%	10%
sample size	20 ▫	40 ▫	80 ▫
SIMPSON	.846	.851	.856
	.003	.001	.000
	.180	.109	.061
	91%	89%	89%
SHANNON	2.03	2.07	2.09
	.047	.016	.007
	.663	.448	.272
	85%	87%	87%

key

PE	average of point estimates
MSE	observed mean square error
CI Length	average of confidence interval lengths
% COV	observed confidence interval coverage

(from 100 replications)

Bootstrap Estimators for a Function of Proportions under Cluster Sampling with Two or More Stages

The developments in the previous sections can be generalized to designs with two or more stages of clustering, where the last stage subclusters are completely observed. Results for a two-stage cluster design are discussed in this section. The previous notation is extended for a two-stage cluster design.

Consider a population subdivided into N_1 clusters, each one subdivided into N_2 subclusters, with M_{ij} elements in the j^{th} subcluster of the i^{th} cluster. Let Y_{ijk} be the k^{th} element of the j^{th} subcluster of the i^{th} cluster, k=1,...,M_{ij}, j=1,...,N_2 and i=1,...,N_1. The population mean by element,

$$\begin{aligned}\overline{\overline{\overline{Y}}} &= \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{M_{ij}} Y_{ijk}}{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} M_{ij}} \\ &= \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{M_{ij}} Y_{ijk} \,/\, N_1N_2}{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} M_{ij} \,/\, N_1N_2},\end{aligned}$$

is a ratio of population means by subclusters. If n_1 clusters are sampled and n_2 subclusters are sampled from each selected cluster, an estimator for the population mean by element is

$$\begin{aligned}\overline{\overline{\overline{y}}} &= \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{m_{ij}} y_{ijk}}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}} \\ &= \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{m_{ij}} y_{ijk} \,/\, n_1n_2}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij} \,/\, n_1n_2},\end{aligned} \tag{3.29}$$

where y_{ijk} is the value for the k^{th} element of the j^{th} sampled subcluster of the i^{th} sampled cluster, and m_{ij} is the value of M_{ij} for the j^{th} sampled subcluster of the

i^{th} sampled cluster, k=1,...,m_{ij}, j=1,...,n_2 and i=1,...,n_1. Equation (3.29) describes the estimator $\bar{\bar{\bar{y}}}$ as a ratio of estimators for population means by subcluster.

The application of bootstrap methods to two-stage cluster sampling creates bootstrap samples by a two-stage procedure. In the first stage, n_1 clusters are randomly selected, with replacement, from the original sample of n_1 clusters. In the second stage, from each resampled cluster, n_2 subclusters are randomly selected, with replacement, from the set of originally sampled subclusters. Thus, a bootstrap sample contains $n_1 \times n_2$ subclusters which is the same number of sampled subclusters in the original sample.

A bootstrap estimator for $f(\bar{\bar{\bar{Y}}})$, based on $f(\bar{\bar{\bar{y}}})$, requires the calculation of the following quantities, from each bootstrap sample:

$$\bar{\bar{\bar{y}}}^{**} = \begin{cases} \dfrac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{m_{ij}^{**}} y_{ijk}^{**}}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**}} & \text{if } \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**} > 0 \\ 0 & \text{if } \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**} = 0, \end{cases} \tag{3.30}$$

and

$$f^{**} = f(\bar{\bar{\bar{y}}}^{**}),$$

where y_{ijk}^{**} is the value for the k^{th} element of the j^{th} resampled subcluster from the i^{th} resampled cluster, and m_{ijk}^{**} is the value of M_{ij} for the j^{th} resampled subcluster from the i^{th} resampled cluster, k=1,...,m_{ij}^{**}, j=1,...,n_2 and i=1,...,n_1.

The bootstrap estimator for $f(\bar{\bar{\bar{Y}}})$ is

$$f_{\text{boot}}^{**} = \frac{1}{B}\sum_{j=1}^{B} f^{**j},$$

where B is the number of bootstrap samples. The bootstrap estimator for the variance of $f(\overline{\overline{y}})$ is

$$Var_{**}(f^{**}) = \frac{1}{B-1}\sum_{j=1}^{B}(f^{**j} - f^{**}_{\text{boot}})^2.$$

The estimator (3.30), from a bootstrap sample, can be expressed as a ratio,

$$\overline{\overline{y}}^{**} = \begin{cases} \dfrac{\dfrac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{m_{ij}^{**}} y_{ijk}^{**}}{n_1 n_2}}{\dfrac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**}}{n_1 n_2}} = \dfrac{\overline{\overline{y}}^{**}}{\overline{\overline{m}}^{**}} & \text{if } \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**} > 0 \\ 0 & \text{if } \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}^{**} = 0. \end{cases}$$

The variance of $\overline{\overline{y}}^{**}$, under bootstrap sampling scheme is

$$Var_{**}(\overline{\overline{y}}^{**}) = \frac{n_1 - 1}{n_1}\frac{s_1^2}{n_1} + \frac{1}{n_1 n_2}\frac{n_2 - 1}{n_2}s_2^2,$$

and according to Cochran (1977, page 278),

$$\widehat{Var}(\overline{\overline{y}}) = \left(1 - \frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + \frac{n_1}{N_1}\left(1 - \frac{n_2}{N_2}\right)\frac{s_2^2}{n_1 n_2},$$

where $s_1^2 = \sum_{i=1}^{n_1}(\overline{y}_i - \overline{\overline{y}})^2/(n_1 - 1)$ and $s_2^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\overline{y}_{ij} - \overline{y}_i)^2/[n_1(n_2 - 1)]$. Therefore, $Var_{**}(\overline{\overline{y}}^{**})$ differs from $\widehat{Var}(\overline{\overline{y}})$.

The same technique as for the one-stage cluster design is used to define a modified bootstrap estimator for $f(\overline{\overline{Y}})$. From each bootstrap sample calculate the following quantities:

$$\begin{aligned} \widetilde{\overline{\overline{y}}}^{**} &= \overline{\overline{y}} + K_1(\overline{\overline{y}}^{*} - \overline{\overline{y}}) + K_2(\overline{\overline{y}}^{**} - \overline{\overline{y}}^{*}), \\ \widetilde{\overline{\overline{m}}}^{**} &= \overline{\overline{m}} + K_1(\overline{\overline{m}}^{*} - \overline{\overline{m}}) + K_2(\overline{\overline{m}}^{**} - \overline{\overline{m}}^{*}), \\ \widetilde{\overline{\overline{y}}}^{**} &= \widetilde{\overline{\overline{y}}}^{**} / \widetilde{\overline{\overline{m}}}^{**}, \end{aligned}$$

and

$$\tilde{f}^{**} = f(\tilde{\bar{\bar{y}}}^{**}),$$

where

$$K_1 = \{[(N_1 - n_1)/N_1][n_{1b}/(n_1 - 1)]\}^{1/2},$$

$$K_2 = \{[(N_2 - n_2)/N_2][n_{1b}/N_1][n_{2b}/(n_2 - 1)]\}^{1/2},$$

n_{1b} and n_{2b} are the number of selections of clusters and subclusters in a bootstrap sample, respectively,

$$\bar{\bar{y}} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{m_{ij}} y_{ijk} \,/\, n_1 n_2,$$

$$\bar{\bar{y}}^{*} = \sum_{i=1}^{n_{1b}}\sum_{j=1}^{n_2}\sum_{k=1}^{m^{*}_{ij}} y^{*}_{ijk} \,/\, n_{1b} n_2,$$

$$\bar{\bar{y}}^{**} = \sum_{i=1}^{n_{1b}}\sum_{j=1}^{n_{2b}}\sum_{k=1}^{m^{**}_{ij}} y^{**}_{ijk} \,/\, n_1 b n_{2b},$$

and y^{*}_{ijk} is the value for the k^{th} element of the j^{th} original subcluster from the i^{th} resampled cluster, and m^{*}_{ij} is the value of M_{ij} for the j^{th} subcluster from the i^{th} resampled cluster, k=1,...,m^{*}_{ij}, j=1,...,n_2 and i=1,...,n_{1b}.

Since $\bar{\bar{y}}$, $\bar{\bar{y}}^{*}$, and $\bar{\bar{y}}^{**}$ are nonnegative, $\bar{\bar{y}}^{**} = (1 - K_1)\bar{\bar{y}} + (K_1 - K_2)\bar{\bar{y}}^{*} + K_2\bar{\bar{y}}^{**}$ is positive if: (i)$0 < 1 - K_1 < 1$, (ii)$0 < K_1 - K_2 < 1$ and (iii)$0 < K_2 < 1$. Condition (i) is satisfied if

$$n_{1b} < (n_1 - 1)N_1/(N_1 - n_1).$$

Condition (iii) is satisfied if

$$n_{2b} < (n_2 - 1)(N_1/n_{1b})N_2/(N_2 - n_2).$$

The right hand side of condition (ii) is satisfied if conditions (i) and (iii) hold. The left hand side of condition (ii) is satisfied if $n_{1b} = n_1 - 1$, $n_{2b} = n_2 - 1$ and

$$[(N_1 - n1)/N_1] < [(N_2 - n2)/N_2][(n_1 - 1)/N_1].$$

If the three conditions are met, it can be proved that $\tilde{\bar{y}}^{**}$ is a value in the interval [0,1], by following the same steps as in the proof of Result 1.

Analogously, for more than two stages of subsampling, a modified bootstrap estimator for a function of a population mean can be defined.

Literature Cited

Cochran, W. G. (1977). Sampling Techniques. John Wiley, New York.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, 37, 36–48

Fuller, W. A. (1976). Introduction to Statistical Time Series. John Wiley, New York.

Heltshe, J. F. and Forrester, N. E. (1983). Estimating Species Richness Using the Jackknife Procedure. Biometrics, 39, 1–11

Heltshe, J. F. and Forrester, N. E. (1985). Statistical Evaluation of the Jackknife Estimate of Diversity when Using Quadrat Samples. Ecology, 66, 107–111

Miller, R. G. (1974). The Jackknife – a Review. Biometrica, 61, 1–15

Pielou, E. C. (1975). Ecological Diversity. John Wiley, New York.

Rao, J. N. K. and Wu, C. F. J. (1985). Inference from Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics. Journal of the American Statistical Association, 80, 620–630

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241

CHAPTER 4. ESTIMATION OF DIVERSITY OF BIRD COMMUNITIES IN FIVE HABITATS

Introduction

Different environments within a city define habitats more or less favorable to animal life. An observational study of the city of Ames, Iowa, is currently being conducted by James Dinsmore, Georgia Bryan and Bret Giesler with the objective of analyzing the structure of bird communities in five habitats:

- commercial areas,
- new residential areas,
- old residential areas,
- green belts,
- parks.

The commercial areas typically have few or no trees, for example, commercial buildings, parking lots, etc. The main difference between the two types of residential areas is that trees are more developed and have overlapping branches in the old residential areas. Green belts are characterized by natural and diverse composition of vegetation. Parks are open areas with dispersed vegetation, subject to greater human

interference. Examples of parks are football fields, playgrounds and green areas with picnic tables.

Results of a statistical analysis for the data obtained during the winter season of 1989-1990 are presented here. The number of species of birds, Simpson index of diversity,

$$D = 1 - \sum_{i=1}^{S} \pi_i^2,$$

and Shannon index of diversity,

$$H = -\sum_{i=1}^{S} \pi_i log(\pi_i),$$

are used to quantify the diversity of species within each of the five habitats, where $\pi_1, ..., \pi_S$ denote the relative abundance of S bird species in a habitat. Inferences are based on bootstrap methods of estimation.

In a concise description, bootstrap methods are computer intensive methods that resample the original sample, a large number of times. The estimates of interest are calculated from each of the created samples, called *bootstrap samples*, and these estimates are used in the calculation of point estimates and confidence intervals. One option is to use the $(1 - \frac{\alpha}{2})100^{th}$ and $(\frac{\alpha}{2})100^{th}$ percentiles of the set of bootstrap estimates as a $(1 - \alpha)100\%$ confidence interval for the parameter of interest. Basic information on bootstrap methods is found in Efron (1982) and Efron and Gong (1983).

The simulation studies in Chapter 2 showed good results for the bootstrap estimation of the number of species, based on a random sample of quadrats, provided that most species are observed in the sample. By repeatedly taking samples of quadrats from the original sample of quadrats, the bootstrap procedure estimates the number

of species that might be missing from the original sample. Consequently, an estimate for the number of species in the habitat is obtained by adding the estimate of the number of missing species to the observed number of species.

Rao and Wu (1988) consider properties of resampling methods, particularly bootstrap estimation, for estimating functions of population means. Further results for one and two-stage cluster sampling were presented in Chapter 3, where it was shown that a bias correction can substantially improve the bootstrap estimation of the Shannon and Simpson indices, from quadrat samples.

The next section presents a brief description of the sampling design used in the bird study. The data are used to illustrate the application of the bootstrap procedure. Estimates of number of species and the Shannon and Simpson indices are evaluated for each of the five habitats.

Nonparametric methods have previously been used in ecological studies to compare diversity measures. Morton and Davidson (1989), for example, employed the Wilcoxon (Mann-Whitney) test to compare diversity indices for communities of harvester ants. The Wilcoxon test (see Hollander and Wolfe, 1973, chapter 4) is a test that compares two populations. The null hypothesis is that both populations have the same continuous distribution, and the alternative hypothesis is that one distribution is exactly as the other, except that it has a shift of location. Therefore, the test is appropriate for comparing populations with the same dispersion. The test is based on the sum of the ranks of the individual random samples, when the combined sample is sorted. Tables are available with critical points for small samples, and normal approximation is used for larger sample sizes. The Wilcoxon test is equivalent to the Mann-Whitney test (in the case of no ties in the ranking of the samples). The

bootstrap method is observed to be more powerful than the Wilcoxon method for finding differences between habitats in the bird study.

Sampling Design and Bootstrap Sampling

For each habitat type, nonoverlapping areas of approximately ten hectares were located on maps: four old residential areas, eight commercial areas, nine green belt areas, ten new residential areas and twelve park areas. The general procedure was to randomly select, without replacement, eight large areas, from each habitat type, and to choose three observation sites, circles with radius of $25m$, from each one. For the old residential areas, where there were only four areas of ten hectares, all areas were surveyed by selecting six observation sites from each one. Therefore, a total of twenty-four sites were chosen from each habitat. An observation site is referred to here as a quadrat. The centers of three quadrats for a particular area were selected by the following procedure:

- the first point is chosen at random,
- a second point is chosen by moving one hundred meters away from the first point, in a randomly chosen direction,
- a third point is chosen by moving one hundred meters away from the second point, at a randomly chosen direction.

For the old residential areas this procedure was repeated twice, independently, but quadrats were not allowed to overlap. In this observational study, it was important to survey all quadrats during the same time period, in order to have the same environmental conditions such as temperature, wind speed and direction, percentage

of cloud cover, etc. The distance of one hundred meters separating quadrats is a practical compromise that allowed for some randomness in the selection of quadrats and also allowed all quadrats in all habitats to be examined in a reasonable short period of time. In each quadrat, records were made on all birds seen or heard during eight-minute intervals in the morning (during the first two hours after sunrise, approximately). Each quadrat was surveyed three times (December 1989, January 1990 and February 1990).

The statistical analysis treats the quadrats as three independent selections from the ten-hectare area chosen at the first stage, although this represents an approximation to the actual sampling scheme. The information obtained in the three different months was pooled for each quadrat.

The application of bootstrap methods to this problem creates bootstrap samples by a two-stage procedure. In the first stage, eight large areas are randomly selected, with replacement, from the set of eight large areas originally selected. In the second stage, three quadrats are randomly selected, with replacement, from the set of originally sampled quadrats, from each resampled large area. For the old residential habitat four random selections, with replacement are made in the first stage, and six quadrats are randomly selected, with replacement, in the second stage. As in the original sample, a bootstrap sample contains twenty-four quadrats. One thousand bootstrap samples were created from the original sample for each habitat.

In the estimation of the number of species, the number of species in the original sample that were omitted in the bootstrap sample was recorded for each of the one thousand bootstrap samples. Point estimates and 90% confidence intervals were obtained for the number of missing species from those one thousand values. Point

estimates and 90% confidence intervals for the number of species were obtained by adding the estimates for the number of missing species to the observed number of species.

In the estimation of the Shannon index, an estimate from the original sample, $\widehat{H}$, is obtained by evaluating the index at the estimated species proportions (number of individuals of a species divided by the total number of individuals, in the pool of all observation sites). This estimate is biased. Estimates for the bias are obtained, from each bootstrap sample by $H^* - \widehat{H}$, where H^* is the index value evaluated at the estimated species proportions from the bootstrap sample. The bias corrected bootstrap estimator for the Shannon index is obtained by subtracting from $\widehat{H}$ the median of one thousand estimates of bias. A 90% confidence interval for the bias was obtained by the 5^{th} and 95^{th} percentiles of the set of estimates of bias. A 90% confidence interval for Shannon index was obtained by subtracting from $\widehat{H}$ the limits of the confidence interval for the bias. The Simpson index was estimated in a similar manner.

Estimation Results

The number of individuals per species observed in each sampled quadrat, for each habitat, during the winter season of 1989-1990, is presented in Appendix C. Histograms of one thousand bootstrap estimates of missing species, for the five habitats studied, are presented in Figures 4.1 and 4.2. Since some of those histograms are skewed to the right, confidence intervals were obtained based on percentiles of the set of bootstrap estimates rather than on large sample normal approximations. The point estimator used was the median of the set of one thousand estimates. An

estimate for the number of species in a habitat was obtained by adding the median of the one thousand bootstrap values for the number of missing species to the observed number of species in the original sample. Table 4.1 presents the observed number of species and point estimates and 90% confidence intervals for the number of species for each habitat. Figure 4.3 provides a graphical comparison of the confidence intervals. The habitats are ordered along the horizontal axis with respect to the observed number of species. Point estimates are indicated by dots inside the 90% confidence intervals for the number of species. Figure 4.3 shows an increasing trend in the point estimates of number of species of commercial areas, parks, new residential areas, old residential areas and green belts.

Confidence intervals and point estimates for the Shannon and the Simpson indices of diversity are presented in Tables 4.2 and 4.3, respectively. These estimates are displayed graphically in Figure 4.4. The habitats are ordered along the horizontal axis with respect to the observed number of species, and point estimates are indicated by dots inside the 90% confidence intervals for the index. Green belts and parks seem to have similar index values. Both residential area types seem to have the same index value, smaller than parks and green belts. The indices of diversity are larger for more even species abundances. In this study, sparrows were much more abundant than other species, in the residential areas. Commercial areas seem to have lower diversity.

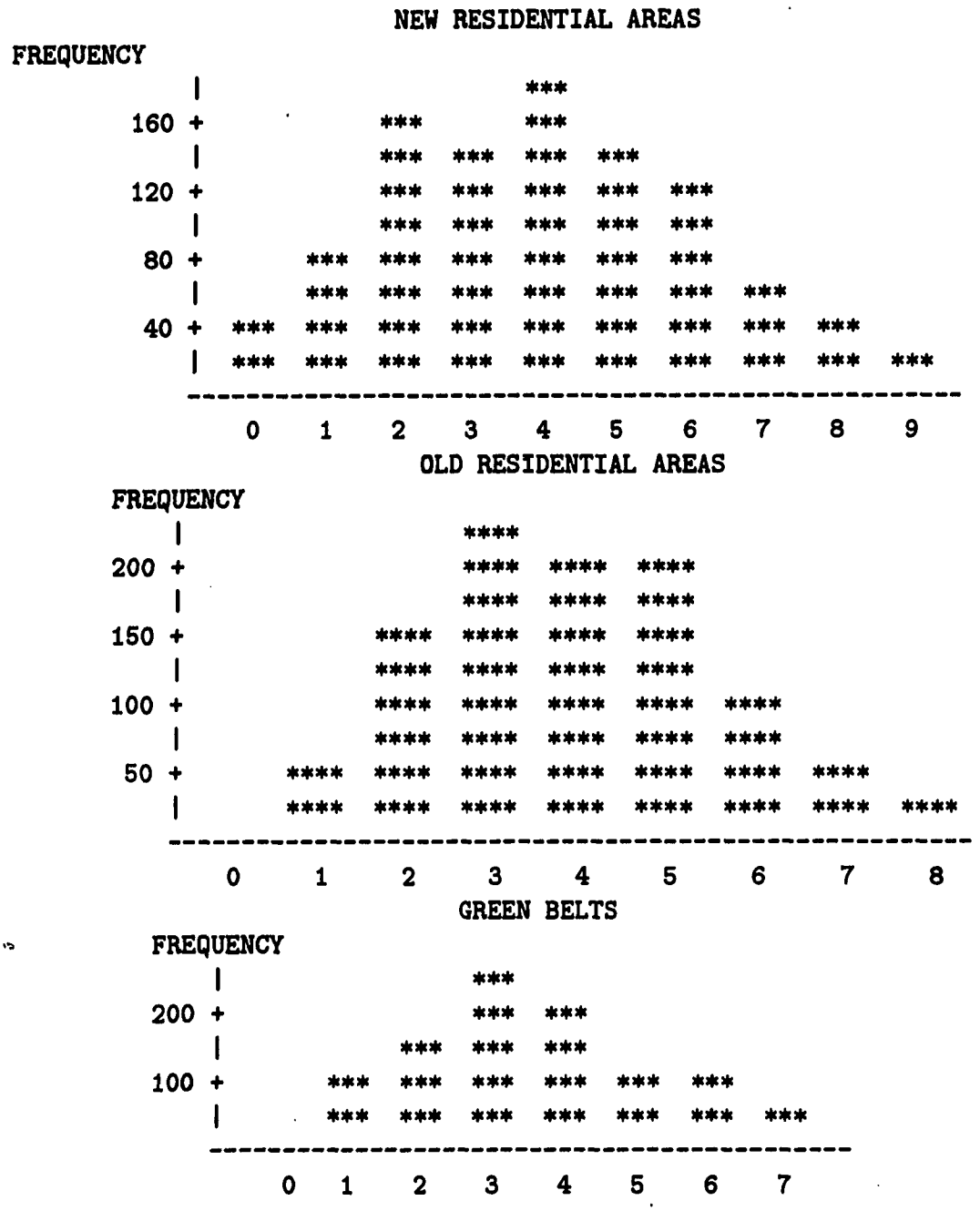


Figure 4.1: Histograms of number of species observed in the original sample, but missing from bootstrap samples (1000 bootstrap samples), for new residential areas, old residential areas and green belts

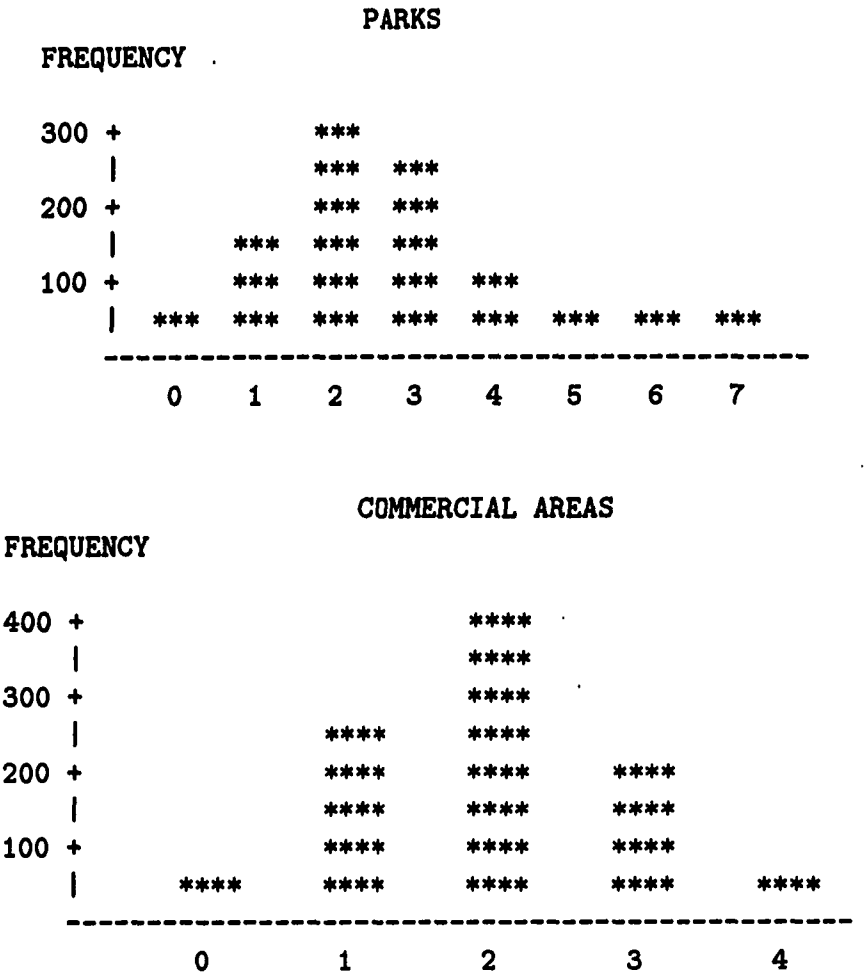


Figure 4.2: Histograms of number of species observed in the original sample, but missing from bootstrap samples (1000 bootstrap samples), for parks and commercial areas

Table 4.1: Observed number of species of birds, bootstrap point estimate and 90% confidence interval for the number of species, for five habitats - commercial areas, parks, new residential areas, old residential areas and green belts in Ames, Iowa (winter 1989-1990)

Habitat	Observed Number of Species	Point Estimate	90% Confidence Interval
commercial	7	9	[8,11]
park	13	16	[14,20]
new residential	16	20	[17,24]
old residential	18	22	[19,25]
green belt	20	23	[21,27]

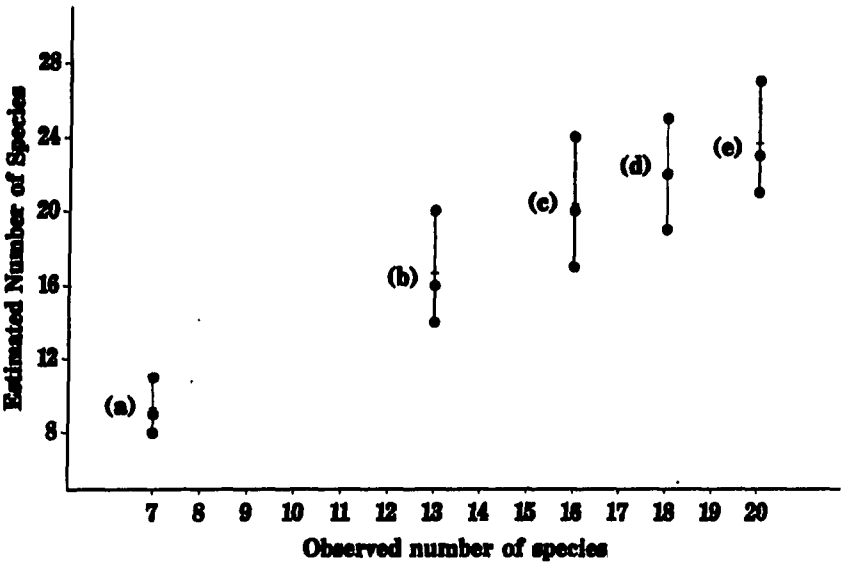


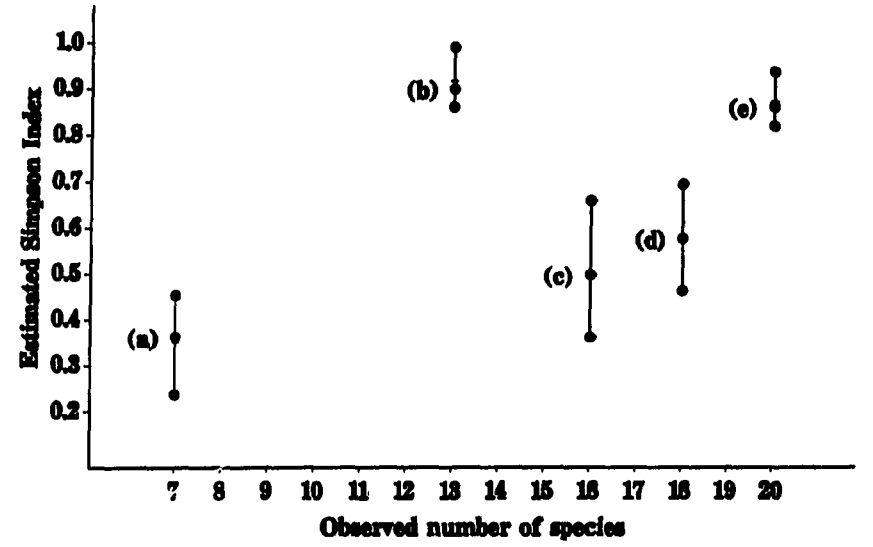
Figure 4.3: Bootstrap point estimate and 90% confidence interval for the number of bird species in five habitats - (a) commercial areas, (b) parks, (c) new residential areas, (d) old residential areas, (e) green belts

Table 4.2: Observed value (from the original sample), bootstrap point estimate and 90% confidence interval for the Shannon index of bird diversity for five habitats - commercial areas, parks, new residential areas, old residential areas and green belts

Bird Habitat	Observed Value	Point Estimate	90% Confidence Interval
commercial	0.73	0.76	[0.53,0.94]
park	2.19	2.42	[2.22,2.82]
new residential	1.31	1.40	[1.03,1.80]
old residential	1.36	1.39	[1.07,1.72]
green belt	2.29	2.4	[2.22,2.71]

Table 4.3: Observed value (from the original sample), bootstrap point estimate and 90% confidence interval for the Simpson index of bird diversity for five habitats - commercial areas, parks, new residential areas, old residential areas and green belts

Bird Habitat	Observed Value	Point Estimate	90% Confidence Interval
commercial	0.362	0.362	[0.237,0.453]
park	0.860	0.898	[0.859,0.989]
new residential	0.489	0.496	[0.360,0.657]
old residential	0.576	0.575	[0.461,0.692]
green belt	0.843	0.858	[0.817,0.934]



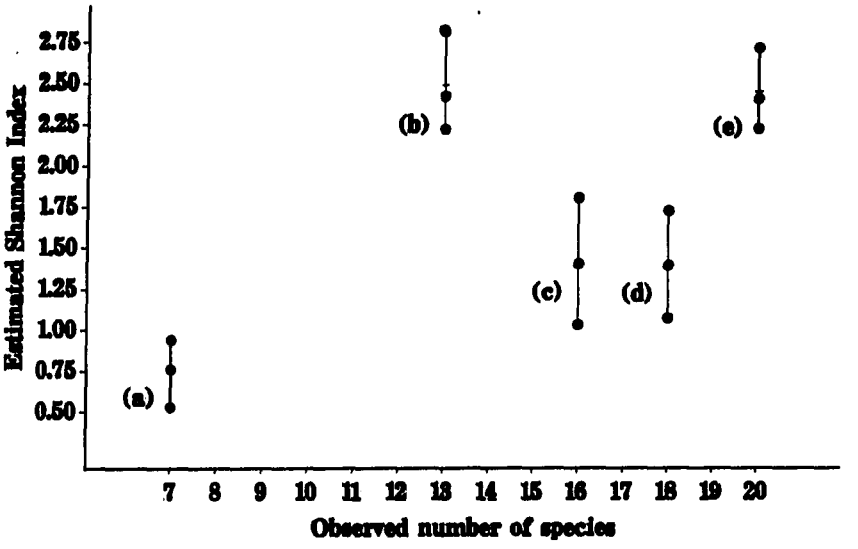


Figure 4.4: Bootstrap point estimates and 90% confidence intervals for the Shannon and Simpson indices of bird diversity in five habitats - (a) commercial areas, (b) parks, (c) new residential areas, (d) old residential areas, (e) green belts

To test whether a diversity index or the number of species in two habitats differ significantly, differences between estimates of the particular diversity measure (diversity index or number of species) for the two habitats were computed for one thousand pairs of bootstrap samples. The 5^{th} and 95^{th} percentiles for the set of one thousand differences were used to define a 90% confidence interval for the difference between the diversity measure in two specific habitats. The inclusion of zero in the interval corresponds to not rejecting the hypothesis of equality of the diversity measure for the two habitats, at a significance level of 0.1. The resulting 90% confidence intervals for the differences between the number of species, Shannon index and Simpson index for each pair of habitats are given in Table 4.4. These intervals indicate that there are fewer species coexisting in commercial areas than in any other habitat type, and that parks accommodate fewer species than old residential areas and green belts. The tests conclusions are the same for both Shannon and Simpson indices, except for the comparison of new residential and commercial areas, where the lower limit of the confidence interval is a negative number for the Simpson index, and a positive number for the Shannon index. The main conclusions are that commercial areas have less diversity than any other habitat type (except for residential areas, if Simpson index is used), and parks and green belts have larger diversity than residential areas. The larger values for the parks and green belts reflect the greater evenness among the the species abundance.

The Wilcoxon test was applied to compare diversity indices for each pair of habitats. Index values were calculated for each $10ha$ area examined in each habitat type by pooling the data for all sampled quadrats in each $10ha$ area. This resulted in eight index values for each habitat type, except old residential areas were only four

index values were calculated. For each pair of habitats, the computed index values were pooled, sorted, and the test statistic was determined by the sums of the ranks of index values for the two habitats. For example, the sampled values of Shannon index for green belts are

1.07, 1.60, 1.61, 1.64, 1.87, 1.94, 2.28, 2.34,

and for new residential areas,

0.12, 0.41, 0.44, 0.85, 1.036, 1.15, 1.50, 1.68.

Pooling and sorting those sets, the sum of the ranks are 95 and 41, for green belts and new residential areas, respectively. Therefore, a difference between Shannon index values is detected, with significance level of 0.002. The results for both methods of comparison are summarized in Table 4.5. The conclusions of the Wilcoxon test coincided with the bootstrap conclusions, for the cases where the bootstrap procedure showed no difference in diversity. However, some differences detected by the bootstrap methods were not detected by the Wilcoxon test. For example, the sampled values of the Shannon index for parks are

0.0, 0.0, 0.96, 1.040, 1.31, 1.81, 1.82, 1.94.

Pooling this set of values with the set of sampled values from new residential areas, the smallest and largest ranks in the sorted set are both from the park habitat. Consequently, the Wilcoxon statistic does not detect a significant shift. The bootstrap method, on the other hand, is able to detect that the mean value of the Shannon index is larger for parks than new residential areas (significance level approximately equal to zero), since most bootstrap samples use information from more than one $10ha$

Table 4.4: 90% confidence intervals for the differences between measures of diversity (number of species, Shannon index and Simpson index) for pairs of habitats

Habitats	Number of Species	Shannon Index	Simpson Index
new residential - old residential	[-7,3]	[-0.486,0.539]	[-0.272,0.122]
new residential - green belt	[-7,0]	[-1.454,-0.633]	[-0.517,-0.203]
new residential - park	[-1,8]	[-1.524,-0.619]	[-0.555,-0.251]
new residential - commercial	[7,15]	[0.178,1.175]	[-0.060,0.366]
old residential - green belt	[-6,3]	[-1.483,-0.668]	[-0.422,-0.160]
old residential - park	[1,10]	[-1.538,-0.647]	[-0.466,-0.200]
old residential - commercial	[10,16]	[0.264,1.002]	[0.073,0.383]
green belt - park	[5,10]	[-0.312,0.230]	[-0.124,0.029]
green belt - commercial	[12,17]	[1.384,2.025]	[0.395,0.647]
park - commercial	[5,10]	[1.387,2.105]	[0.436,0.710]

area. No significant differences were found by the Wilcoxon test for all comparisons involving old residential areas, except for the green belts for the Simpson index. One possible explanation might be the small sample size for the old residential habitat.

The Wilcoxon test was not applied to to compare number of species for a pair of habitats, since the differences between number of species are integers, and many ties appear in a combined sample.

For the estimation of an index of diversity, or for a general function of variable means, bootstrap procedures might present greater power for detecting differences than the nonparametric procedures that have been previously used in such studies. Confidence intervals are obtained through bootstrap procedures based on a simple random sample of quadrats or based on some complex sampling design.

Table 4.5: Significance levels for the bootstrap and Wilcoxon methods for comparing bird diversity (Shannon and Simpson indices) among habitats (the notation "ns" indicates that the difference between the two habitats is not significant at the 0.1 level)

Habitats	Shannon		Simpson	
	Bootstrap	Wilcoxon	Bootstrap	Wilcoxon
new residential-old residential	ns	ns	ns	ns
new residential-green belt	0.000	0.004	0.000	0.002
new residential-park	0.000	ns	0.000	ns
new residential-commercial	0.026	ns	ns	ns
old residential-green belt	0.000	ns	0.000	0.072
old residential-park	0.000	ns	0.000	ns
old residential-commercial	0.002	ns	0.004	ns
green belt-park	ns	ns	ns	ns
green belt-commercial	0.000	0.002	0.000	0.014
park - commercial	0.000	ns	0.000	ns

Literature Cited

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS 38, SIAM–NSF

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, 37, 36–48

Hollander, M. and Wolfe, D. A. (1973). Nonparametric Statistical Methods. John Wiley, New York

Morton, S. R. and Davidson, D. W. (1988) Comparative Structure of Harvester Ant Communities in Arid Australia and North America. Ecological Monographs, 58(1), 19–38

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241

CHAPTER 5. SUMMARY

This dissertation focus on applications of resampling methods to the estimation of measures of diversity. Species richness and indices of diversity are studied separately.

Simulation studies and a real data problem showed that the resampling estimators are less accurate when there are many rare species in a community. Conditional on additional knowledge that most species were observed, the variability among quadrats will determine an adequate confidence interval for the number of species when bootstrap estimators are used.

Bootstrap methods are reliable to obtain confidence intervals for a diversity index, when cluster sampling is employed.

LITERATURE CITED

Bickel, P. J, and Freedman, D. A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. The Annals of Statistics, 12, 470–482

Cochran, W. G. (1977). Sampling Techniques. John Wiley, New York.

Colinvaux, P. (1978). Why Big Fierce Animals Are Rare, an Ecologist's Perspective. Princeton University Press, New Jersey

Colinvaux, P. (1986). Ecology. John Wiley, New York.

Crawley, M. J. (1986). Plant Ecology. Blackwell Scientific Publications, London

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS 38, SIAM–NSF

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, 37, 36–48

Fauth, J. E., Crother, B. I. and Slowinski, J. B. (1989). Elevational Patterns of Species Richness, Evenness, and Abundance of the Costa Rican Leaf-Litter Herpetofauna. Biotropica, 21(2), 178–185

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The Relation between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. Journal of Animal Ecology, 12, 42–58

Fuller, W. A. (1976). Introduction to Statistical Time Series. John Wiley, New York.

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, 40, 237–264

Heltshe, J. F. and Forrester, N. E. (1983). Estimating Species Richness Using the Jackknife Procedure. Biometrics, 39, 1–11

Heltshe, J. F. and Forrester, N. E. (1985). Statistical Evaluation of the Jackknife Estimate of Diversity When Using Quadrat Samples. Ecology, 66, 107–111

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: a Critique and Alternative Parameters. Ecology, 52, 577–586

Krebs, C. J. (1985). Ecology: The Experimental Analysis of Distribution and Abundance (3rd ed.). Harper & Row, New York.

Krebs, C. J. (1989). Ecological Methodology.

Miller, R. G. (1974). The Jackknife – a Review. Biometrica, 61, 1–15

Mingoti, S. A. (1989). Estimating the Total Number of Distinct Species When Quadrat Sampling or Sampling by Elements Is Used. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa

Minshall, G. W., Petersen, R. C. and Nimz, C. F. (1985). Species Richness in Streams of Different Size from the Same Drainage Basin. The American Naturalist, 125, 16–38

Morton, S. R. and Davidson, D. W. (1988) Comparative Structure of Harvester Ant Communities in Arid Australia and North America. Ecological Monographs, 58(1), 19–38

Nilsson, S. G., Bengtsson, J. and As, S. (1988). Habitat Diversity or Area Per Se? Species Richness of Woody Plants, Carabid Beetles and Land Snails on Islands. Journal of Animal Ecology, 57, 685–704

Patil, G. P. and Taillie, C. (1982). Diversity as a Concept and its Measurement. Journal of the American Statistical Association, 77, 548–561

Peet, R. K. (1974). The Measurement of Species Diversity. Annual Review of Ecology and Systematics, 5, 285–307

Pielou, E. C. (1975). Ecological Diversity. John Wiley, New York.

Pielou, E. C. (1977). Mathematical Ecology. John Wiley, New York.

Rao, J. N. K. and Wu, C. F. J. (1985). Inference from Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics. Journal of the American Statistical Association, 80, 620–630

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241

Ross, S. T., Matthews, W. J. and Echelle, A. A. (1985). Persistence of Stream Fish Assemblages: Effects of Environmental Changes. The American Naturalist, 126, 24–40

Shannon, C. E. and Weaver W. (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana

Simpson, E. H. (1949). Measurement of Diversity. Nature, 163, 688

Smith, E. P. and van Belle, G. (1984). Nonparametric Estimation of Species Richness. Biometrics, 40, 119–129

Slobodkin, L. B. and Fishelson, L. (1974). The Effect of the Cleaner-Fish Labroides Dimidiatos on the Point Diversity of Fishes on the Reef Front at Eilat. American Naturalist, 108, 369–376

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Kenneth Koehler, for all support and advice directed to the development of this dissertation.

I would like to thank the members of my committee, Dr. Yasuo Amemiya, Dr. William Clark, Dr. David Harville and Dr. Paul Hinz, for their important comments and suggestions of improvements.

I am particularly indebted to Dr. William Clark for his valuable advice and help on ecological aspects of my research.

APPENDIX A. COMPUTER PROGRAM FOR GENERATION OF COMMUNITY

```
c       This program generates (in a unit square)
c       coordinates representing individuals
c       belonging to "s" species.
c
c       An input file, "values.dta", is needed.
c       The first line must have three values:
c       s=number of species,
c       dseed= first seed for the number generator, and
c       sigma=standard deviation of normal distribution.
c
c       A total of "s" lines follow, with two numbers
c       in each line:
c       plam=Poisson parameter for # of parents, and
c       olam=Poisson parameter for # of offspring.
c
c       The output includes: the coordinates of all
c       individuals of the community;
c       a line with the total number of individuals
c       for a species preceding lines that contain
c       the coordinates of the individuals for that
c       species (each line corresponds to an
c       individual).
c
c       (ggpon is the IMSL Poisson generator)
c       (ggubs is the IMSL uniform generator)
c       (ggnpm is the IMSL normal generator)
```

```
      real off(2), u(2)
      real*8 dseed
      open(unit=8,file='pop.dta',status='unknown')
      open(unit=7,file='values.dta',status='unknown')
      read(7,*)s,dseed,sigma
      isoma=0
      do 2 j=1,s
      read(7,*)plam,olam
      call ggpon (plam,dseed,1,np,ier)
      call ggpon (olam,dseed,1,no,ier)
      ntotal=np+np*no
      write(8,*)ntotal
      isoma=isoma+ntotal
      do 3 i=1,np
      call ggubs (dseed,2,u)
      write (8,*) u(1),u(2)
      do 4 k=1,no
      call ggnpm (dseed,1,r)
      call ggubs (dseed,1,t)
      theta=3.1415927*t
      dist=sigma*r
      off(1)=u(1)+dist*cos(theta)
      off(2)=u(2)+dist*sin(theta)
      write(8,*) off(1),off(2)
4     continue
3     continue
2     continue
      end
```

APPENDIX B. RESULTS FOR THE PROOF OF CONSISTENCY OF A BOOTSTRAP ESTIMATOR FOR THE VARIANCE OF $f(\overline{y})$

For the following development s, t, w and v are integers in the set $\{1,\cdots,S\}$.

Let

$$d_s = \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\overline{y}_s^* - \overline{y}_s).$$

Since each bootstrap sample is selected independently using simple random sampling with replacement,

$$\begin{aligned} E_*(\overline{y}_s^*) &= E_*\left(\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{j=1}^{m_i^*} y_{ijs}^*\right) \\ &= \frac{1}{n_b}\sum_{i=1}^{n_b}\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{m_k} y_{kjs} \\ &= \overline{y}_s, \end{aligned} \tag{B.1}$$

and,

$$E_*(d_s) = 0.$$

The variance of d_s under the bootstrap sampling scheme is

$$\begin{aligned} Var_*(d_s) &= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right) Var_*(\overline{y}_s^*) \\ &= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)\frac{n-1}{n_b}\,\frac{s_s^2}{n} \end{aligned} \tag{B.2}$$

$$= \left(\frac{N-n}{N}\right) \frac{s_s^2}{n}$$

$$= O(1)\, O(n^{-1})\, O_p(1) \tag{B.3}$$

$$= O_p(n^{-1}). \tag{B.4}$$

Result (B.3) follows (3.10), noting that the bootstrap sample size is n_b; therefore the fraction $[(n-1)/n_b]$ is used instead of $[(n-1)/n]$. Result (B.4) follows since, under the original sample scheme, s_s^2 is a random variable with expected value S_s^2, considered to be bounded.

The expected value of the product $d_s d_t$ under the bootstrap sampling scheme is

$$\begin{aligned}
E_*(d_s d_t) &= E_*\left[\left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\overline{y}_s^* - \overline{y}_s)\left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\overline{y}_t^* - \overline{y}_t)\right] \\
&= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right) E_*[(\overline{y}_s^* - \overline{y}_s)(\overline{y}_t^* - \overline{y}_t)] \\
&= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right) E_*\left[\left(\frac{\sum_{i=1}^{n_b}\sum_{j=1}^{m_i^*} y_{ijs}^*}{n_b} - \overline{y}_s\right)\right. \\
&\quad \left.\times\left(\frac{\sum_{i=1}^{n_b}\sum_{j=1}^{m_i^*} y_{ijt}^*}{n_b} - \overline{y}_t\right)\right] \\
&= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right) E_*\left[\frac{\sum_{i=1}^{n_b}(y_{is}^* - \overline{y}_s)}{n_b}\;\frac{\sum_{i=1}^{n_b}(y_{it}^* - \overline{y}_t)}{n_b}\right] \\
&= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\,\frac{1}{n_b}\right) E_*\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{j=1}^{n_b}[\,(y_{is}^* - \overline{y}_s)\,(y_{jt}^* - \overline{y}_t)\,]\right\} \\
&= \left(\frac{N-n}{N}\,\frac{1}{n-1}\right) E_*\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}[\,(y_{is}^* - \overline{y}_s)\,(y_{it}^* - \overline{y}_t)\,]\right\}. \qquad \text{(B.5)}
\end{aligned}$$

Equation (B.6) follows since, for $i \neq j$,

$$E_*[(y_{is}^* - \overline{y}_s)\,(y_{jt}^* - \overline{y}_t)] = E_*(y_{is}^* - \overline{y}_s)E_*(y_{jt}^* - \overline{y}_t) = 0,$$

by independence of bootstrap selections (with replacement), and

$$\begin{aligned} E_*(y^*_{is}) &= E_* \left(\sum_{j=1}^{m^*_i} y^*_{ijs} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ijs} \\ &= \overline{y}_s. \end{aligned}$$

Therefore,

$$\begin{aligned} E_*(d_s d_t) &= \left(\frac{N-n}{N} \frac{1}{n-1} \right) E_* \left\{ \frac{1}{n_b} \sum_{i=1}^{n_b} [\, (y^*_{is} - \overline{y}_s)\, (y^*_{it} - \overline{y}_t)\,] \right\} \\ &= \left(\frac{N-n}{N} \frac{1}{n-1} \right) \frac{1}{n} \sum_{i=1}^{n} [\, (y_{is} - \overline{y}_s)\, (y_{it} - \overline{y}_t)\,] \\ &= \left(\frac{N-n}{N} \frac{1}{n} \right) \left[\frac{1}{n-1} \sum_{i=1}^{n} [\, (y_{is} - \overline{y}_s)\, (y_{it} - \overline{y}_t)\,] \right] \\ &= \frac{N-n}{N} \frac{1}{n} s_{st} \qquad \text{(B.6)} \\ &= \widehat{Cov}(\overline{y}_s, \overline{y}_t). \qquad \text{(B.7)} \end{aligned}$$

From equation (B.8), the order of probability of $E_*(d_s d_t)$ is established,

$$\begin{aligned} E_*(d_s d_t) &= \frac{N-n}{N} \frac{1}{n} s_{st} \\ &= O(1)\, O(n^{-1})\, O_p(1) \\ &= O_p(n^{-1}), \qquad \text{(B.8)} \end{aligned}$$

where, $n_b^{-1} = O(n^{-1})$, and by the boundedness of S_{st}.

The expected value of the product $d_s d_t d_w$ under the bootstrap sampling scheme is

$$E_*(d_s d_t d_w) = E_* \left[\left(\frac{N-n}{N} \frac{n_b}{n-1} \right)^{1/2} (\overline{y}^*_s - \overline{y}_s) \right.$$

$$\left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\bar{y}_t^* - \bar{y}_t)$$
$$\left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{1/2}(\bar{y}_w^* - \bar{y}_w)\Big]$$

$$= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{3/2} E_*\left[(\bar{y}_s^* - \bar{y}_s)\right.$$
$$\left.(\bar{y}_t^* - \bar{y}_t)(\bar{y}_w^* - \bar{y}_w)\right]$$

$$= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{3/2} \frac{1}{n_b^3} E_*\left\{\sum_{i=1}^{n_b}\sum_{j=1}^{n_b}\sum_{k=1}^{n_b}\right.$$
$$\left.\left[(y_{is}^* - \bar{y}_s)\,(y_{jt}^* - \bar{y}_t)\,(y_{kw}^* - \bar{y}_w)\right]\right\}$$

$$= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{3/2} \frac{1}{n_b^2} E_*\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\Big[\,(y_{is}^* - \bar{y}_s)\right.$$
$$\left.(y_{it}^* - \bar{y}_t)\,(y_{iw}^* - \bar{y}_w)\,\Big]\right\}$$

$$= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{3/2} \frac{1}{n_b^2}\,\frac{1}{n}\sum_{i=1}^{n}\Big[\,(y_{is} - \bar{y}_s)$$
$$(y_{it} - \bar{y}_t)\,(y_{iw} - \bar{y}_w)\,\Big]$$

$$= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{3/2} \frac{1}{n_b^2}\,\frac{n-1}{n}\,s_{\mathsf{stw}}$$

$$= \left(\frac{N-n}{N}\right)^{3/2}\left[\frac{1}{n_b(n-1)}\right]^{1/2}\frac{1}{n}\,s_{\mathsf{stw}}$$

$$= O(1)\,O(n^{-1})\,O(n^{-1})\,O_p(1) \tag{B.9}$$

$$= O_p(n^{-2}), \tag{B.10}$$

where (B.11) follows the boundedness of S_{stw} and the assumption $n_b = O(n^{-1})$.

Similarly,

$$E_*(d_s d_t d_w d_v) = \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^{2} \frac{1}{n_b^4} E_*\left\{\sum_{i=1}^{n_b}\sum_{j=1}^{n_b}\sum_{k=1}^{n_b}\sum_{l=1}^{n_b}\right.$$
$$\left.\left[(y_{is}^* - \bar{y}_s)\,(y_{jt}^* - \bar{y}_t)\,(y_{kw}^* - \bar{y}_w)\,(y_{lv}^* - \bar{y}_v)\right]\right\}$$

$$
\begin{aligned}
&= \left(\frac{N-n}{N}\,\frac{n_b}{n-1}\right)^2 \frac{1}{n_b^3} E_* \left\{ \frac{1}{n_b} \sum_{i=1}^{n_b} \right. \\
&\quad \left. \left[(y_{is}^* - \overline{y}_s)\,(y_{it}^* - \overline{y}_t)\,(y_{iw}^* - \overline{y}_w)\,(y_{iv}^* - \overline{y}_v) \right] \right\} \\
&= \left(\frac{N-n}{N}\right)^2 \left[\frac{1}{n_b(n-1)(n)}\right] s_{stwv} \\
&= O_p(n^{-3}). \qquad \text{(B.11)}
\end{aligned}
$$

APPENDIX C. BIRD HABITAT DATA

Bird species observed in **Green Belts:**

1. House Sparrow
2. Starling
3. Chickadee
4. Blue Jay
5. White-breasted Nuthatch
6. Crow
7. Cardinal
8. Robin
9. Goldfinch
10. Downy Woodpecker
11. Rock Dove
12. Belted Kingfisher
13. Red-bellied Woodpecker
14. Pine Siskin
15. Brown Creeper
16. Cedar Waxwing

17. Dark-eyed Junco
18. Red-breasted Nuthatch
19. Tufted Titmouse
20. Owl

The following matrix presents the number of individuals per species from **Green Belts**, where the columns represent the 20 species labelled above and the rows represent quadrats. Rows 1 to 3, 4 to 6, ...,22 to 24 correspond to quadrats selected from the same $10ha$ area.

```
 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0
 1 3 4 0 3 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0

 0 0 7 0 1 0 2 0 0 0 0 0 0 0 0 0 1 0 0 0
 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 5 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0

 0 0 6 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3 0 8 0 1 0 0 0 4 1 0 0 0 5 0 0 4 1 0 1
 0 0 5 0 5 0 0 0 0 0 2 0 0 0 0 1 0 0 0 0

 0 0 4 1 3 0 0 0 1 3 0 0 0 0 0 0 0 0 0 0
 3 0 4 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 2 0
18 1 5 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0

 0 3 6 3 2 0 2 1 1 1 0 0 0 3 0 0 0 0 0 0
 0 0 0 0 3 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0
 2 0 5 0 5 1 2 0 7 0 0 0 0 7 0 0 2 0 0 0

 0 0 1 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 4 0 2 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0
 1 0 5 0 2 0 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0

 0 0 6 0 0 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0
 0 0 2 1 0 2 0 0 0 1 0 1 0 0 0 0 0 0 0 0
 0 0 7 0 3 0 0 0 0 3 0 1 0 0 0 2 0 0 0 0

 0 0 0 6 0 0 2 0 0 1 0 0 0 0 0 0 5 0 0 0
 7 0 3 0 1 0 0 0 1 0 2 0 0 0 0 0 0 0 0 0
 0 0 4 0 0 2 0 0 0 4 0 0 0 0 0 0 0 0 0 0
```

Bird species observed in **Old Residential** areas:

1. House Sparrow
2. Starling
3. Chickadee
4. Blue Jay
5. White-breasted Nuthatch
6. Crow
7. Cardinal
8. Goldfinch
9. Downy Woodpecker
10. Rock Dove
11. Hairy Woodpecker
12. Pine Siskin
13. Brown Creeper
14. Dark-eyed Junco
15. Red-headed Woodpecker
16. House Finch
17. Red-breasted Nuthatch
18. Tufted Titmouse

The following matrix presents the number of individuals per species from **Old Residential** areas, where the columns represent the 18 observed species labelled above and the rows represent quadrats. Rows 1 to 6, 7 to 12, 13 to 18, and 19 to 24 correspond to quadrats selected from the same $10ha$ area.

38	10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
26	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	3	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
7	3	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	10	1	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0
25	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
17	7	6	0	3	0	1	0	1	0	1	0	0	1	0	0	0	0
7	2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
15	1	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	1
7	7	1	3	0	0	0	0	2	0	0	0	0	0	0	0	0	0
12	0	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
7	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	10	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0	0
9	2	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
5	1	0	1	3	0	0	0	1	0	0	0	0	0	0	0	0	0
8	4	2	2	0	0	6	4	0	0	0	0	0	2	0	0	0	0
12	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	0	0	0	0	9	0	0	1	1	0	0	0	0	0	0	0
12	3	4	3	2	0	5	0	0	0	0	2	0	0	0	0	2	0
30	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Bird species observed in **New Residential** areas:

1. House Sparrow
2. Starling
3. Chickadee
4. Blue Jay
5. White-breasted Nuthatch
6. Crow
7. Cardinal
8. Goldfinch
9. Downy Woodpecker
10. Rock Dove
11. Pine Siskin
12. Brown Creeper
13. Cedar Waxwing
14. Dark-eyed Junco
15. Field Sparrow
16. House Finch

The following matrix presents the number of individuals per species from **New Residential** areas, where the columns represent the 16 observed species labelled above and the rows represent quadrats. Rows 1 to 3, 4 to 6, ...,22 to 24 correspond to quadrats selected from the same $10ha$ area.

```
 7 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

14 0 1 0 0 0 0 0 0 0 0 0 2 0 0 0
 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2 0 0 2 0 0 0 0 1 0 0 0 0 0 0 0

11 0 0 0 0 0 0 0 0 0 0 2 0 0 1 0 0
 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

16 3 0 1 0 0 1 0 0 2 1 0 0 0 0 0
10 3 1 0 0 0 1 0 1 3 0 0 0 1 0 0
 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

 4 0 0 1 0 0 2 0 0 0 0 0 7 0 0 0
 1 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0
36 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0

23 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

11 0 0 5 0 0 1 0 0 0 0 0 0 0 2 0
 5 0 3 1 0 0 0 1 1 0 0 0 0 0 0 0
 1 0 0 1 1 0 0 3 0 0 0 1 0 0 0 0

 6 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0
 7 0 0 1 0 0 0 0 0 0 0 0 0 3 0 2
 2 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Bird species observed in **Parks:**

1. House Sparrow
2. Chickadee
3. Blue Jay
4. White-breasted Nuthatch
5. Crow
6. Cardinal
7. Downy Woodpecker
8. Red-bellied Woodpecker
9. Pine Siskin
10. Brown Creeper
11. Dark-eyed Junco
12. Red-breasted Nuthatch
13. Ring-necked Pheasant

The following matrix presents the number of individuals per species from **Parks**, where the columns represent the 13 observed species labelled above and the rows represent quadrats. Rows 1 to 3, 4 to 6, ...,22 to 24 correspond to quadrats selected from the same $10ha$ area.

```
0 2 0 0 0 0 0 0 0 0  0 0 0
0 2 1 0 0 0 0 0 0 0  0 0 4
0 0 0 0 0 0 0 0 0 0  0 0 0

1 0 0 0 0 0 2 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0

0 2 1 0 0 0 1 0 0 1  0 0 0
0 4 0 1 0 0 5 1 0 0  1 1 0
0 0 0 0 0 0 1 0 0 0  0 0 0

0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0

0 1 0 0 0 0 0 0 0 0  5 0 0
4 0 0 1 0 1 0 0 0 0  2 0 0
0 4 3 3 0 1 1 1 0 0 10 0 0

0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0

1 0 0 0 0 0 0 0 0 1  0 0 0
1 0 0 0 0 0 0 0 2 0  0 0 0
1 0 0 0 0 0 0 0 0 0  0 0 0

0 1 1 1 1 0 1 0 0 0  0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0
0 0 0 0 0 0 1 0 0 0  0 0 0
```

Bird species observed in **Commercial** areas:

1. House Sparrow
2. Starling
3. Blue Jay
4. White-breasted Nuthatch
5. Crow
6. Cardinal
7. Rock Dove

The following matrix presents the number of individuals per species from **Commercial** areas, where the columns represent the 7 observed species labelled in the previous page, and the rows represent quadrats. Rows 1 to 3, 4 to 6, ...,22 to 24 correspond to quadrats selected from the same $10ha$ area.

```
 0 0 0 0 0 0 0
 0 0 0 0 0 0 0
 0 3 0 0 0 0 0

 0 0 0 0 1 0 0
 3 1 0 0 0 0 0
 0 0 0 0 0 0 0

 9 0 0 0 0 0 2
 5 0 0 0 0 0 0
 0 0 0 0 0 0 0

 8 1 0 0 0 0 0
 0 0 0 0 0 0 0
13 2 0 0 0 0 0

 0 0 0 0 0 0 0
 0 0 0 1 1 0 0
 0 0 0 0 0 0 0

19 5 0 0 1 1 0
10 3 0 0 0 0 0
 0 0 0 0 0 0 0

 0 0 0 0 0 0 0
17 2 1 0 0 0 0
 1 0 0 0 0 0 0

14 2 0 0 0 0 0
 7 4 0 0 0 0 0
 5 0 0 0 0 0 0
```